

Adaptive MCMC with online relabeling

Extended version

Rémi Bardenet¹, Olivier Cappé², Gersende Fort², and Balázs Kégl^{1,3}

¹*Laboratoire de Recherche en Informatique, Univ. Paris-Sud XI*
e-mail: remi.bardenet@gmail.com

²*LTCI, Telecom ParisTech & CNRS*
e-mail: cappe@telecom-paristech.fr; gfort@telecom-paristech.fr

³*CNRS, Laboratoire de l'Accélérateur Linéaire, Univ. Paris-Sud XI*
e-mail: balazs.kegl@gmail.com

Abstract: When targeting a distribution that is *artificially* invariant under some permutations, Markov chain Monte Carlo (MCMC) algorithms face the *label-switching* problem, rendering marginal inference particularly cumbersome. Such a situation arises, for example, in the Bayesian analysis of finite mixture models. Adaptive MCMC algorithms such as adaptive Metropolis (AM), which self-calibrates its proposal distribution using an online estimate of the covariance matrix of the target, are no exception. To address the label-switching issue, *relabeling* algorithms associate a permutation to each MCMC sample, trying to obtain reasonable marginals. In the case of adaptive Metropolis [15], an *online* relabeling strategy is required. This paper is devoted to the AMOR algorithm, a provably consistent variant of AM that can cope with the label-switching problem. The idea is to nest relabeling steps within the MCMC algorithm based on the estimation of a *single covariance matrix* that is used *both* for adapting the covariance of the proposal distribution in the Metropolis algorithm step *and* for online relabeling. We compare the behavior of AMOR to similar relabeling methods. In the case of compactly supported target distributions, we prove a strong law of large numbers for AMOR and its ergodicity. These are the first results on the consistency of an online relabeling algorithm to our knowledge. The proof underlines latent relations between relabeling and vector quantization.

Keywords and phrases: Adaptive Markov chain Monte Carlo; label-switching; stochastic approximation; vector quantization

AMS 2000 Mathematics Subject Classification: Primary 65C05, 60J10; Secondary 62F15.

Contents

1	Introduction	2
2	The AMOR algorithm	4
	2.1 The algorithm	4
	2.2 An illustrative example	5
3	Convergence results	12
	3.1 A stable AMOR algorithm	12
	3.2 Convergence of stable AMOR	14
4	Conclusion	16
5	Appendix: proofs	17
	5.1 Preliminary results	17
	5.2 Differentiating the cross-entropy term in (3.11)	17

5.3	The Lyapunov function	23
5.4	Proof of Proposition 3.1	25
5.5	Regularity in θ of the Poisson solution	28
5.6	Proof of Theorem 3.2	35
5.7	Proof of Theorem 3.3	40
5.8	Proof of Theorem 3.4	40
	Acknowledgements	42
	References	42

1. Introduction

Markov chain Monte Carlo (MCMC) is a generic approach for exploring complex probability distributions based on sampling [24]. It has become the *de facto* standard tool in many applications of Bayesian inference. However, a very common situation in which MCMC algorithms face serious difficulties is when the target posterior distribution is known to be invariant under some permutations (or block permutations) of the variables. In that case, the difficulties are both computational, as most often the MCMC algorithm fails to validly visit all the modes of the posterior, and inferential, in particular rendering marginal posterior inference about the individual variables particularly cumbersome [10]. In the literature, this latter difficulty is usually referred to as the *label switching problem* [32]. The most well-known example of this situation is when performing Bayesian inference in a mixture model. In this case the mixture likelihood is invariant to permuting the mixture components and, most often, the prior itself does not favor any specific ordering of the mixture components [9, 32, 17, 18, 22, 31, 19]. Another important example arises in signal processing with additive decomposition models. In this case, the observed signal is represented as the superposition of exchangeable signals, and the main goal is to recover the individual signals or their parameters. In addition, often the number of signals also has to be determined [30, 29, 6]. It was observed empirically that when the dimension of the model is not known, the reversible jump sampler [23] makes it easier to visit the multiple modes corresponding to the permutations but, of course, marginal inference becomes harder due to the additional difficulty of associating components between models of varying dimension.

In this contribution, we address the label switching problem in the generic case where no useful external information on the target is known. This corresponds, for instance, to a posterior distribution when neither the likelihood is assumed to have a specific form, nor the prior is chosen to have conjugacy properties, which forbids the use of Gibbs sampling or other specialized sampling strategies. We assume, however, that the target is known to be invariant under some permutations of the parameters. This framework is typical, for instance, in experimental physics applications where the likelihood computation is commonly deferred to a *black-box* numerical code. In those cases, one cannot assume anything about the structure of the posterior or its conditional distributions, except that they should be invariant to some permutations of the parameters. We also restrict ourselves to the case where the dimension of the model is finite and known so the parameters of the model are \mathbb{R}^d -valued for some fixed and finite d .

Adaptive MCMC algorithms can self-calibrate their internal parameters along the iterations in order to reach decent performance without (or with almost no) knowledge about the target distribution, eliminating the grueling step of tuning the proposals. Adaptive MCMC has been an active field of research in the last ten years, following the pioneering contribution of [15] — see [3] as well as the other papers in the same special issue of *Statistics and Computing*, along

with [4, 2, 28]. Adaptive Metropolis (hereafter AM; [15]) and its variants aim at identifying the unknown covariance structure of the target distribution along the run of a random walk Metropolis-Hastings algorithm with a multivariate Gaussian proposal. The rationale behind this approach is based on scaling results which suggest that, when d tends to $+\infty$, the chain correlation is minimized when the covariance matrix used in the proposal distribution matches, up to a constant that depends on the dimension, the covariance matrix of the target, for a large class of unimodal target distributions with independent marginals [25, 26]. AM thus progressively adapts, using a stochastic approximation scheme, the covariance of the proposal distribution to the estimated covariance of the target.

It has been empirically observed in [5], and we provide further evidence of this fact below in Section 2.2, that the efficiency of AM can be greatly impaired when label switching occurs. The reason for such a difficulty is obvious: if label switching occurs, the estimated covariance matrix no longer corresponds to the local shape of the modes of the posterior and so the exploration can be far from optimal. In Section 2.2, we also provide some empirical evidence that off-the-shelf solutions to the label-switching problem, such as imposing identifiability constraints or post-processing the simulated sample, are not fully satisfactory. A key difficulty here is that most of the approaches proposed in the literature are based on post-processing of the simulated trajectories *after* the MCMC algorithm has been fully run [32, 17, 18, 22, 31, 19, 30]. Unfortunately, in the case of adaptive MCMC, post-processing cannot solve the improper exploration issue described above. On the other hand, online relabeling algorithms [23, 10, 11] often require manual tuning based on, for example, prior knowledge on the location of the redundant modes of the target. Without such manual tuning they often yield poor samplers, as we will show it in Section 2.2.

Our main purpose in this paper is to provide a provably consistent variant of AM that can cope with the label-switching problem. In [5], we proposed an adaptive Metropolis algorithm with online relabeling, called AMOR, based on the original idea of [9]. The idea is to nest relabeling steps within the MCMC algorithm based on the estimation of a *single covariance matrix* that is used *both* for adapting the covariance of the proposal distribution used in the Metropolis algorithm step *and* for online relabeling. Contrary to [9], the AMOR algorithm also corrects for the relabelings using a modified acceptance ratio.

In Section 2.2, we provide empirical evidence that the coupling established in AMOR between the criterion used for relabeling and the estimation of the covariance of the local modes of the posterior is beneficial to avoid the distortion of the marginal distributions. Furthermore, the example considered in Section 2.2 also demonstrates that the AMOR algorithm samples from non-trivial identifiable restrictions of the posterior distribution, that is, truncations of the posterior on regions where the posterior marginals are distinct but from which the complete posterior can be recovered by permutation. The study of the convergence of AMOR in Section 3 reveals an interesting connection with the problem of optimal probabilistic quantization [13] which was implicit in earlier works on label switching. It was observed previously by [21] that some adjustments to the usual theory of stochastic approximation are necessary to analyze online optimal quantification due to the presence of points where the mean field of the algorithm is not differentiable. To circumvent this difficulty, we introduce the stable AMOR algorithm, a novel variant of the AMOR algorithm that avoids these problematic points of the parameter space. Finally, we establish consistency results for the stable AMOR algorithm, showing that it indeed asymptotically provides samples distributed under a suitably defined restriction of the posterior distribution in which the parameters are marginally identifiable.

The paper is organized as follows. In Section 2, we describe the AMOR algorithm and compare it with alternative approaches on an illustrative example. In Section 3, we address the convergence of the algorithm. The detailed proofs are provided in Appendix.

2. The AMOR algorithm

In this section, we briefly review the AMOR algorithm and illustrate its performance on an artificial example.

2.1. The algorithm

Let π be a density with respect to (w.r.t.) the Lebesgue measure on \mathbb{R}^d which is invariant to the action of a group \mathcal{P} of matrices, that is,

$$\forall x \in \mathbb{R}^d, \forall P \in \mathcal{P}, \pi(x) = \pi(Px) .$$

Denote by \mathcal{C}_d^+ the set of $d \times d$ real positive definite matrices. For $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathcal{C}_d^+$, define $L_\theta : \mathbb{R}^d \rightarrow \mathbb{R}_+$ by

$$L_\theta(x) = (x - \mu)^T \Sigma^{-1} (x - \mu) , \quad (2.1)$$

and let $\mathcal{N}(\cdot | \mu, \Sigma)$ denote the Gaussian density with mean μ and covariance matrix Σ . Algorithm 1 describes the pseudocode of AMOR [5].

Algorithm 1.

```

AMOR( $\pi(\cdot), X_0, T, \theta_0 = (\mu_0, \Sigma_0), c, (\gamma_t)_{t \geq 0}$ )
1    $\mathcal{S} \leftarrow \emptyset$ 
2   for  $t \leftarrow 1$  to  $T$ 
3        $\Sigma \leftarrow c \Sigma_{t-1} \triangleright$  scaled adaptive covariance
4        $\tilde{X} \sim \mathcal{N}(\cdot | X_{t-1}, \Sigma) \triangleright$  proposal
5        $\tilde{P} \sim \arg \min_{P \in \mathcal{P}} L_{\theta_{t-1}}(P\tilde{X}) \triangleright$  pick an optimal permutation
6        $\tilde{X} \leftarrow \tilde{P}\tilde{X} \triangleright$  permute the proposal
7       if  $\frac{\pi(\tilde{X}) \sum_P \mathcal{N}(PX_{t-1} | \tilde{X}, \Sigma)}{\pi(X_{t-1}) \sum_P \mathcal{N}(P\tilde{X} | X_{t-1}, \Sigma)} > \mathcal{U}[0, 1]$  then
8            $X_t \leftarrow \tilde{X} \triangleright$  accept
9       else
10           $X_t \leftarrow X_{t-1} \triangleright$  reject
11           $\mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\} \triangleright$  update the posterior sample
12           $\mu_t \leftarrow \mu_{t-1} + \gamma_t (X_t - \mu_{t-1})$ 
13           $\Sigma_t \leftarrow \Sigma_{t-1} + \gamma_t ((X_t - \mu_{t-1})(X_t - \mu_{t-1})^\top - \Sigma_{t-1})$ 
14           $\theta_t \leftarrow (\mu_t, \Sigma_t)$ .
15  return  $\mathcal{S}$ 

```

To explain the proposal mechanism of AMOR, let μ_{t-1} and Σ_{t-1} denote the sample mean and the sample covariance matrix, respectively, at the end of iteration $t - 1$, and let $\theta_{t-1} =$

$(\mu_{t-1}, \Sigma_{t-1})$. Let also \mathcal{S} denote the MCMC sample at the end of iteration $t - 1$. At iteration t , a point \tilde{X} is first drawn from a Gaussian centered at the previous state X_{t-1} and with covariance $c\Sigma_{t-1}$, where c implements the optimal scaling results in [25, 26] discussed in Section 1 (Steps 3 and 4). Then in Steps 5 and 6, \tilde{X} is replaced by $\tilde{P}\tilde{X}$, where \tilde{P} is a uniform draw over the permutations in $\arg\min_P L_{\theta_{t-1}}(P\tilde{X})$ that minimize the relabeling criterion (2.1)¹. This relabeling step makes the augmented sample $S \cup \{\tilde{P}\tilde{X}\}$ look as Gaussian as possible among all augmented sets $S \cup \{P\tilde{X}\}$, $P \in \mathcal{P}$. Formally, it can be seen as a projection onto the Voronoi cell $V_{\theta_{t-1}}$, where

$$V_{\theta} = \{x \in \mathbb{X} / L_{\theta}(x) \leq L_{\theta}(Px), \forall P \in \mathcal{P}\}. \quad (2.2)$$

Then, in Steps 7 to 10, the candidate $\tilde{P}\tilde{X}$ is accepted or rejected according to the usual Metropolis-Hastings rule. Finally, the sample mean and covariance are adapted according to a stochastic approximation scheme in Steps 12 to 14 and so (γ_t) is a sequence of nonnegative steps, usually set according to a polynomial decay $\gamma_t \sim t^{-\beta}$, $\beta \in (1/2, 1]$.

AMOR is a doubly adaptive MCMC algorithm since it is adaptive both in its *proposal* and *relabeling* mechanisms. This means that, besides the proposal distribution, its target also changes with the number of iterations. In Section 3 we will prove that, at each iteration t , AMOR implements a random walk Metropolis-Hastings kernel with stationary distribution $\pi_{\theta} \propto \pi \mathbb{1}_{V_{\theta}}$.

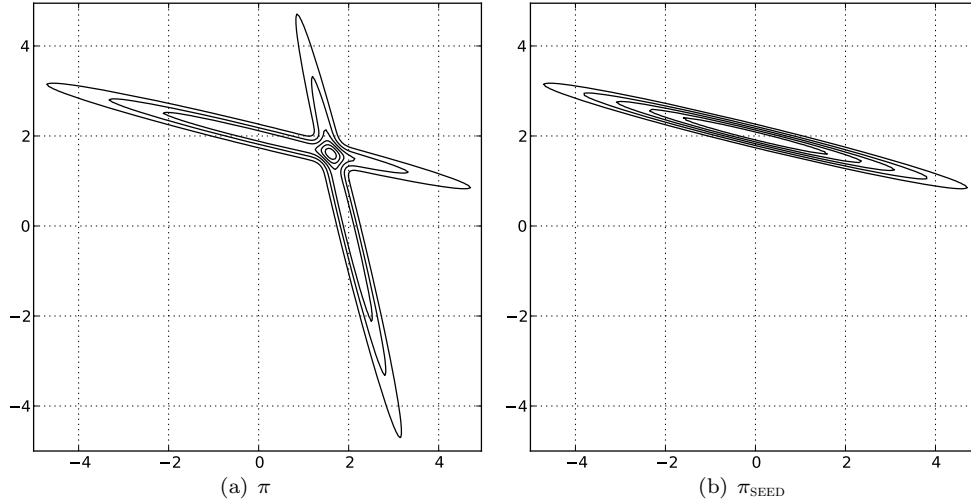


FIG 1. Panel 1(a) shows the target distribution π used in Section 2.2, obtained by symmetrizing the Gaussian π_{SEED} shown in Panel 1(b). π_{SEED} has mean $(0, 2)$ and covariance matrix with diagonal $(16, 1)$ and non-diagonal terms equal to -0.975 .

2.2. An illustrative example

In this section, we consider an artificial target aimed at illustrating the gap in performance between the AMOR algorithm and other common approaches to the label switching problem,

¹Step 5 usually boils down to selecting the permutation \tilde{P} that minimizes $L_{\theta_{t-1}}$. In case of ties, however, \tilde{P} should be drawn uniformly over the set on which the minimum is achieved.

which are compatible with adaptive MCMC. Consider the two-dimensional pdf π depicted in Figure 1(a), which satisfies $\pi(x) = \pi(Px)$ for $P \in \mathcal{P}$, where

$$\mathcal{P} = \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \right\}.$$

The density π is a mixture of two densities with equal weights obtained by superposing the Gaussian pdf π_{SEED} represented in Figure 1(b) with a symmetrized version of itself. This artificial target does not correspond to the posterior distribution in an actual inference problem. In particular, although π itself is a mixture, it is not the posterior distribution of the parameters of any specific mixture model. Nevertheless, it is relevant because it is permutation invariant and the desired solution of the label switching problem is well-defined: we know that, under suitable relabeling, we can obtain univariate near-Gaussian marginals for both coordinates by recovering the marginals of the two-dimensional Gaussian π_{SEED} in Figure 1(b). In spite of its simplicity, this example is challenging because the two marginals of π_{SEED} have similar means (0 and 2) and one has large variance, which makes them hard to separate. Given the modest dimension of the problem, we fix the number of MCMC iterations to 20 000, of which 4 000 are discarded as burn-in. For each algorithm, we assess the quality of the relabeling strategy by looking at the corresponding restriction π' of the target π , and we assess the efficiency of the sampling by plotting the autocorrelation function of each sample and comparing the sample histograms with the marginals of π' .

The results obtained when applying AM, without any relabeling, are shown in Figure 2. The marginal posteriors are sampled quite well (Figures 2(c) and 2(d)) and the covariance of the joint sample (indicated by a thick ellipse Figure 2(a)) is almost symmetric. This is not surprising: the joint distribution, although severely non-Gaussian, is unimodal, and the number of iterations is large enough for AM to explore both the original seed π_{SEED} and its symmetric version by frequent label switching. On the other hand, the covariance of the joint distribution π (Figure 1(a)) is broader than the covariance of the seed π_{SEED} (Figure 1(b)). This results in poor adaptive proposals and slow mixing as indicated by the slight differences between the marginals and the sample marginals, and by the autocorrelation function of the first component of the sample in Figure 2(b). The reference (dashed line) is the autocorrelation function of an MCMC chain with optimal covariance (proportional to the covariance of the target) targeting the single Gaussian π_{SEED} (Figure 1(b)).

We now consider a modified version of AM with online relabeling obtained by simply ordering the variables, meaning that after each proposal $x = (x_1, x_2)$, the components of the proposed point are permuted so that $x_1 \leq x_2$. This strategy is known as *imposing an identifiability constraint*. It is known to perform badly when the constraint does not respect the topology of the target [19]. The results of this approach on our illustrative example are shown in Figure 3. The unshaded triangle in Figure 3 shows that this time the sample is restricted to a subregion of \mathbb{R}^2 where the components are identifiable. Unfortunately, the marginals of π restricted to the unshaded triangle in Figures 3(c) and 3(d) are even more highly skewed than the marginals of the full joint distribution π . In addition, sampling from the restricted distribution π' is not easier than before indicated by the autocorrelation function in Figure 3(b).

Applying the ordering constraint *after* the full sample has been drawn with AM leads to similar results as shown in Figure 4. This shows that the problem lies with the relabeling criterion rather than with the online nature of the relabeling procedure.

Next, we consider the approach introduced by Celeux in [9]. Celeux's algorithm builds on

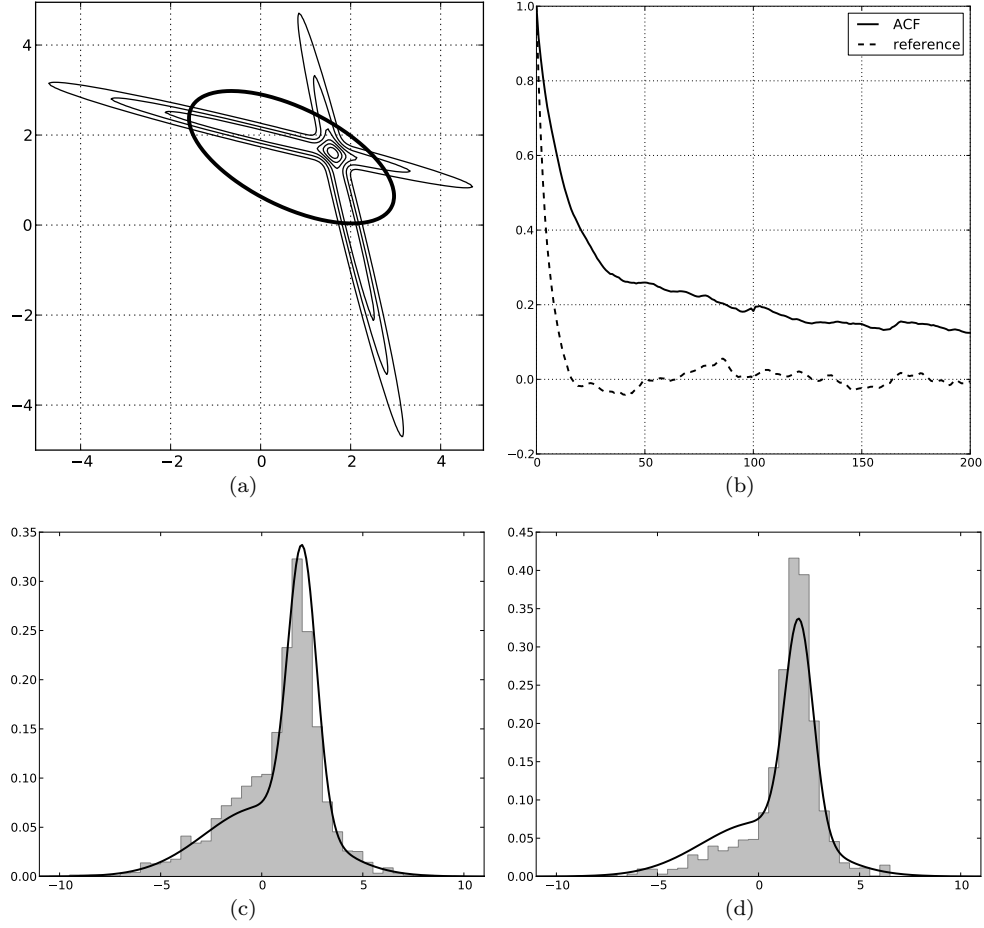


FIG 2. Results of vanilla AM on the two-dimensional target π of Figure 1. The rest of the caption is the same for Figures 3 to 6. On Panel 2(a), level lines of π are depicted in thin black lines; a thick ellipse centered at the empirical mean μ_T of the sample S indicates the set $\{x : (x - \mu_T)^T \Sigma_T^{-1} (x - \mu_T) = 1\}$, where Σ_T is the sample covariance. When appropriate, the region of the space selected by (the last iteration of) the algorithm corresponds to the unshaded background while the region not selected is shaded. On Panel 2(b), the autocorrelation function (ACF) of the first component of S is plotted as a solid line. The dashed line indicates the ACF obtained when sampling from the seed Gaussian π_{SEED} of Figure 1(b) using a random walk Metropolis algorithm with an optimally tuned covariance matrix. Panels 2(c) and 2(d) display the histograms of the two marginal samples. The solid curves are the marginals of π in this figure. In Figures 3 to 6, they are the marginals of π restricted to the unshaded region selected by the algorithms.

a non-adaptive random-walk Metropolis, where online relabeling is performed in the following way: when a point $x = (x^{(1)}, x^{(2)})$ is proposed at time t , it is relabeled by

$$x \leftarrow \arg \min \left\{ \begin{pmatrix} x^{(1)} - \mu_t^{(1)} \\ x^{(2)} - \mu_t^{(2)} \end{pmatrix}^T D_t^{-1} \begin{pmatrix} x^{(1)} - \mu_t^{(1)} \\ x^{(2)} - \mu_t^{(2)} \end{pmatrix}, \begin{pmatrix} x^{(2)} - \mu_t^{(1)} \\ x^{(1)} - \mu_t^{(2)} \end{pmatrix}^T D_t^{-1} \begin{pmatrix} x^{(2)} - \mu_t^{(1)} \\ x^{(1)} - \mu_t^{(2)} \end{pmatrix} \right\}. \quad (2.3)$$

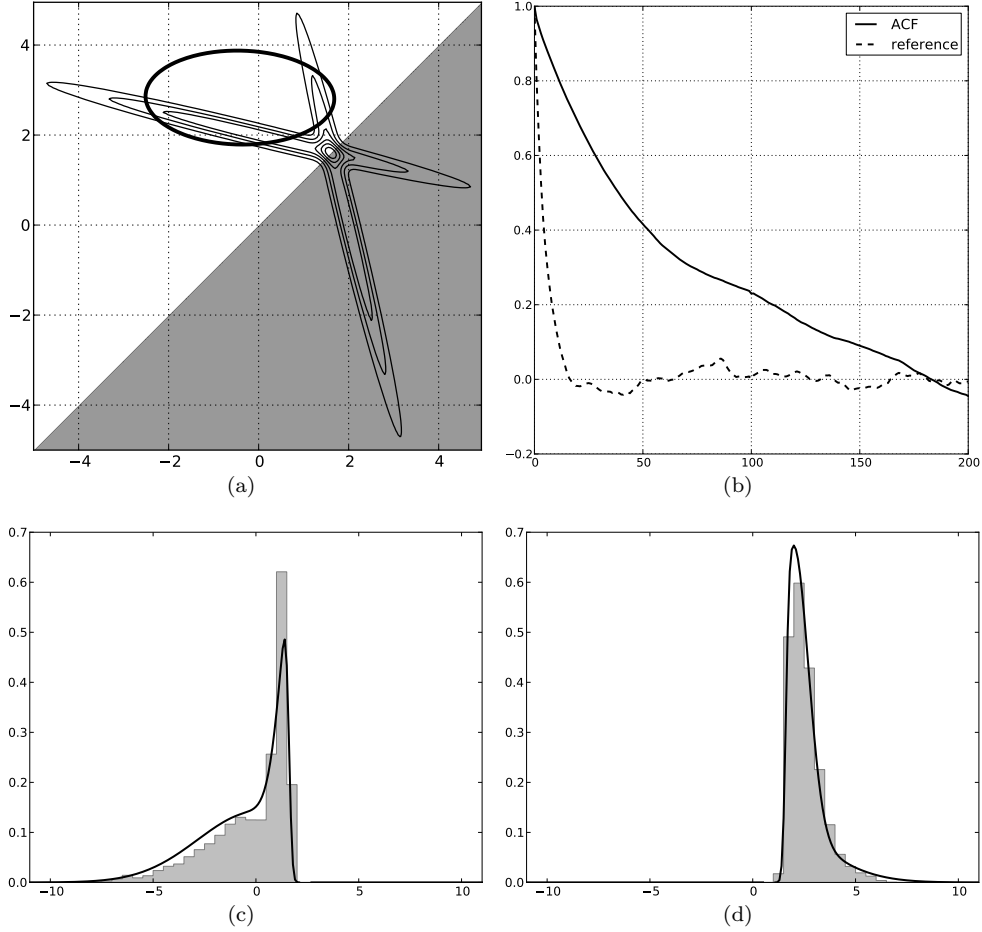


FIG 3. Results of AM with online ordering constraint. For details about the plots, see the caption of Figure 2.

where $\mu_t = (\mu_t^{(1)}, \mu_t^{(2)})$ is the empirical mean of the current sample $x_{1:t} = x_1, \dots, x_t$ and D_t is the diagonal matrix containing the empirical variances of the coordinates of $x_{1:t}$ on its diagonal. Formally, this relabeling rule is equivalent to Steps 6 and 7 of Algorithm 1, but with all non-diagonal elements of Σ equal to zero. The results of Celeux's algorithm are shown in Figure 5. It is hard to determine precisely the formal target of the algorithm. In particular, given the non-isotropic shape of the target, we used a non-isotropic Gaussian proposal with diagonal covariance matrix, and while the preservation of the detailed balance condition then requires incorporating a term into the acceptance ratio to account for the relabeling, it is absent in this approach. It is still possible that the algorithm is *approximately* sampling from the restriction π' of π to this unshaded area in Figure 5 (which represents the relabeling rule implemented at the end of the run) in a certain sense. The histograms in Figures 5(c) and 5(d) are in agreement with the solid line marginals. Certainly, there are no formal guarantees that this should happen. On the other hand, in Section 3 we can prove the corresponding claim for the AMOR algorithm.

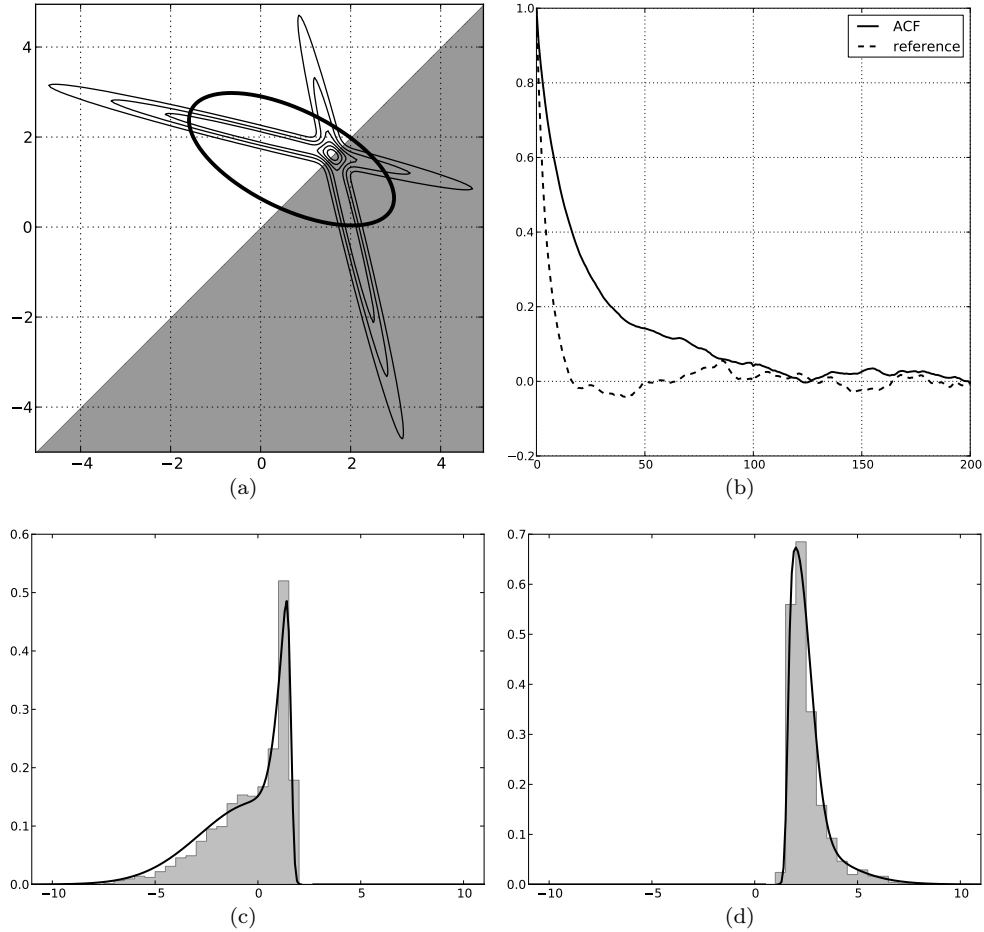


FIG 4. Results of AM with ordering constraint applied as post-processing. For details about the plots, see the caption of Figure 2.

This relabeling strategy seems to recover π_{SEED} better than the mere ordering of coordinates as suggested by the marginal plots in Figures 5(c) and 5(d) which are less skewed and now roughly centered at the correct values (0 and 2, respectively). However, using a diagonal covariance D_t also generates some distortion which results in a severely non-Gaussian, bimodal marginal in Figure 5(c). Because of these imperfections and due to the uncorrelated proposal, the autocorrelation in Figure 5(b) indicates, again, a much less efficient sampling than in the case of an optimal Metropolis chain targeting π_{SEED} .

The significance of Celeux’s algorithm is that its adaptive relabeling rule (2.3) makes it possible to resolve the permutation invariance problem in a non-trivial way which appears to be more adapted to the true geometry of the target. It is still not perfect, and, as suggested by [32], one should replace the diagonal covariance matrix in (2.3) by the full covariance matrix of the sample. However, [32] explored this idea only as a post-processing approach. A severe

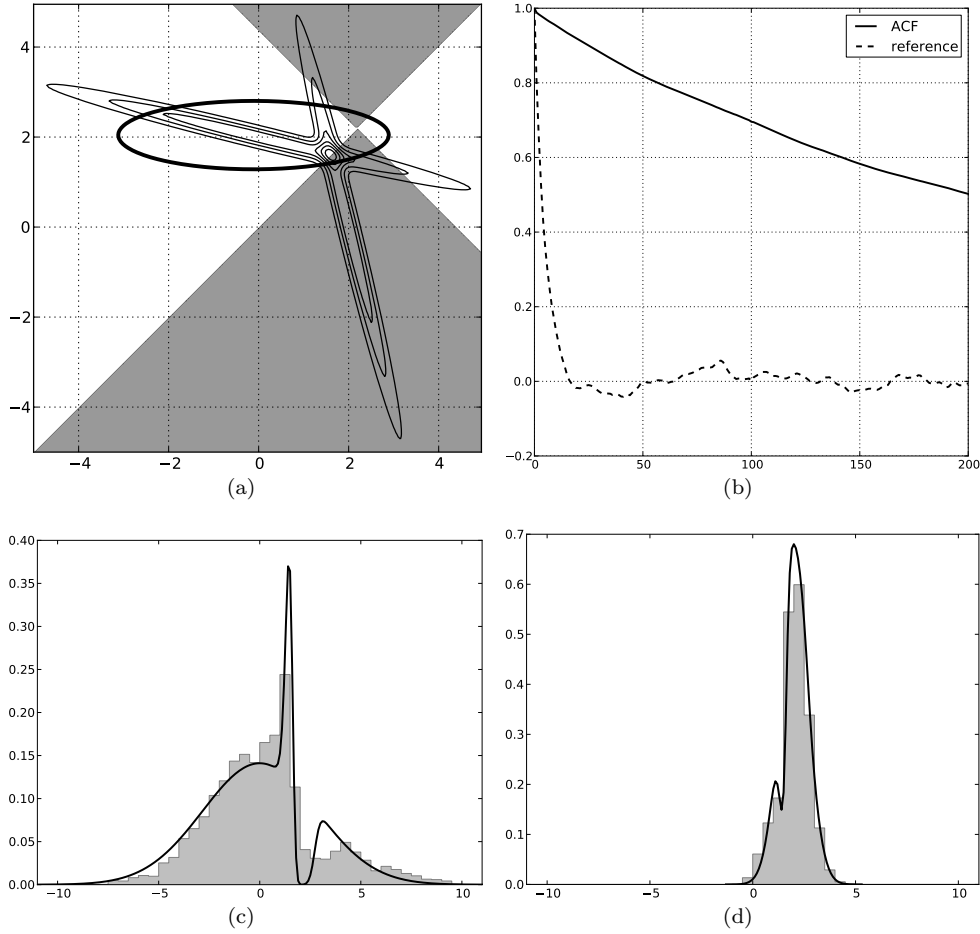


FIG 5. Results of Celeux's algorithm. For details about the plots, see the caption of Figure 2.

difficulty in this context is the computational cost: if T denotes the number of drawn samples and p is the number of permutations to which π is invariant, the required post-processing is a combinatorial problem with p^T possible relabelings. This eventually led [32] to consider a more tractable alternative instead. More importantly in our context, we have seen above (e.g., in Figure 2) that running an adaptive MCMC on the full permutation-invariant target may result in a poor mixing performance. To achieve both relevant relabeling and efficient adaptivity, the key idea of AMOR is to link the covariance of the proposal distribution and the covariance used for relabeling, which are proportional to each other in AMOR.

Figure 6 displays the results obtained using AMOR on our running example. AMOR does separate \mathbb{R}^2 in two regions that respect the topology of the target much more closely than the approaches examined previously. Figure 6(a) indicates that the relabeled target is as Gaussian as possible among all partitionings based on a quadratic criterion of the form (2.1). The marginal histograms in Figures 6(c) and 6(d) now look almost Gaussian. They closely match the

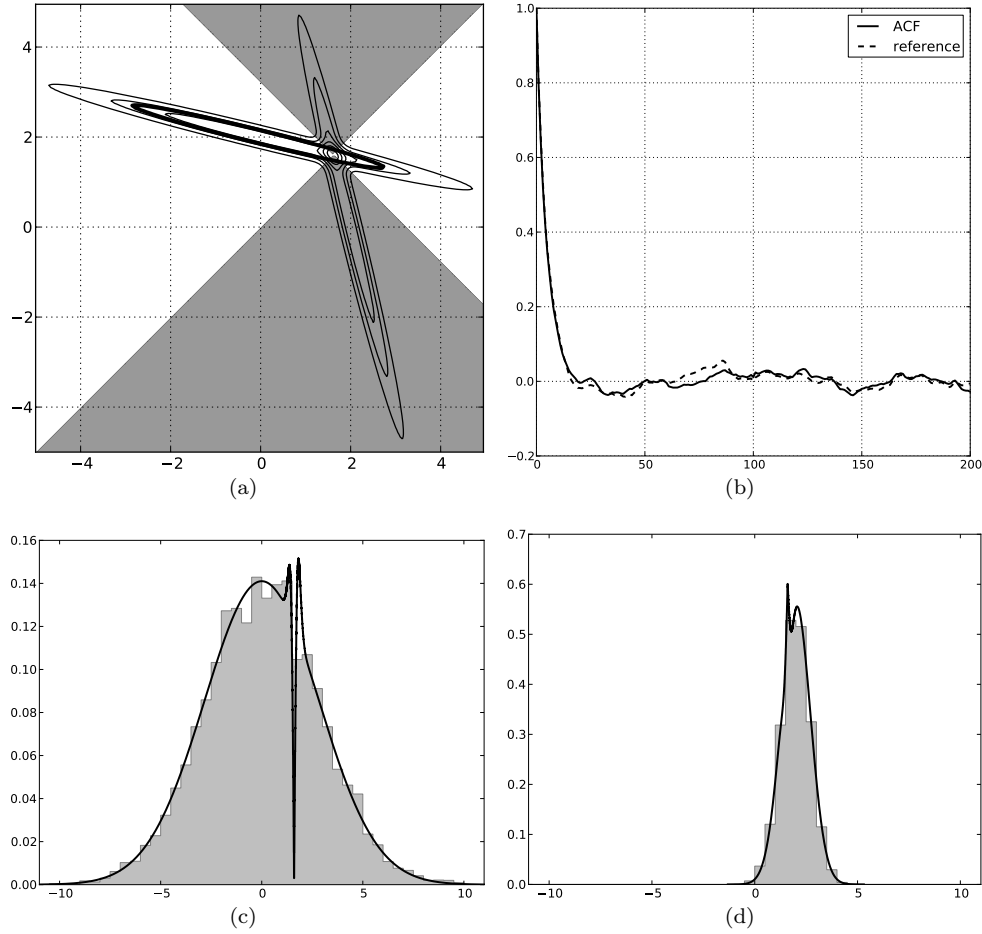


FIG 6. Results of AMOR. For details about the plots, see the caption of Figure 2.

marginals of both the restricted distribution π' and the seed distribution π_{SEED} in Figure 1(b). Furthermore, the autocorrelation function of AMOR (Figure 6(b)) is as good as the reference autocorrelation function corresponding to an optimally tuned random walk Metropolis-Hastings algorithm targeting the seed Gaussian π_{SEED} in Figure 1(b). This perfect adaptation is possible because the sample covariance now matches the covariance of the target restricted to the unshaded region of the plane (Figure 6(a)).

On this example, the AMOR algorithm thus automatically achieves, without any tuning, a satisfactory result that cannot be obtained with any of the methods examined previously. We are now ready to prove our main result which shows that, under suitable conditions, a stable version of AMOR indeed asymptotically samples from the target distribution restricted to a region on which the marginals are identifiable, and that the sample mean and covariance converge to the corresponding moments of the restricted target.

3. Convergence results

AMOR can be cast into the family of adaptive MCMC algorithms in which the updating rule of the design parameter relies on a stochastic approximation scheme. Adaptive MCMC can be described as follows: given a family of transition kernels $(P_\theta)_{\theta \in \Theta}$, the algorithm produces a $(\mathbb{X} \times \Theta)$ -valued process $((X_t, \theta_t))_{t \geq 0}$ such that the conditional distribution of X_t given the past is given by the transition kernel $P_{\theta_{t-1}}$. This algorithm is designed so that when t tends to infinity, the distribution of X_t converges to the invariant distribution of the kernel P_{θ_t} . Convergence of such adaptive procedures was recently analyzed by [27, 12]. In particular, [27] provided sufficient conditions in terms of the so-called containment condition and diminishing adaptation. Furthermore, [12] showed that when each transition kernel P_θ has its own invariant distribution, a condition on the convergence of these distributions is also required.

In Section 3.1, we will show that each transition kernel of AMOR has its own invariant distribution. Therefore, as a preliminary step for the convergence of AMOR, the stability and the convergence of the design parameter sequence $(\theta_t)_{t \geq 0}$ have to be established. Sufficient conditions for the convergence of stochastic approximation procedures rely on the existence of a (sufficiently regular) Lyapunov function on Θ , on the behavior of the mean field at the boundary of the parameter set Θ , and on the magnitude of the stepsize sequence $(\gamma_t)_{t \geq 0}$. For Algorithm 1, we were only able to design a Lyapunov function for which some boundaries of Θ are not repulsive [5]. Therefore, we introduce in this paper a stable AMOR algorithm (Algorithm 2), which differs from Algorithm 1 in the update rules 12 and 13. In particular, we add (i) a penalty in steps 12 and 13 to make the boundaries of Θ repulsive, and (ii) a stabilization step to ensure that the sequence $(\theta_t)_{t \geq 0}$ is bounded.

We prove the convergence of the stable AMOR algorithm under the condition that the support of π is compact.

Assumption 1. π is a density w.r.t. the Lebesgue measure on \mathbb{R}^d , which is bounded and with compact support \mathbb{X} , and which is invariant to permutations in the group \mathcal{P} :

$$\forall x \in \mathbb{X}, \forall P \in \mathcal{P}, \pi(Px) = \pi(x) .$$

The compactity assumption makes it simpler to analyze the limiting behavior of the algorithm. The proofs can be extended to a more general case by using the same tools as in [12] and [1, section 3]. These technical steps are out of the scope of this paper.

This section is organized as follows. In Section 3.1, we first describe the stable AMOR algorithm, and we show that it is an adaptive MCMC algorithm. We then characterize the limiting behavior of the sequence $(\theta_t)_{t \geq 0}$ in Section 3.2 and address a strong law of large numbers for the samples $(X_t)_{t \geq 0}$, as well as the ergodicity of the sampler. All proofs are given in the Appendix.

3.1. A stable AMOR algorithm

Set $\mathcal{P}^* = \mathcal{P} \setminus \{\text{Id}\}$ and

$$\Theta = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathcal{C}_d^+ / \forall P \in \mathcal{P}^*, \Sigma^{-1}\mu \neq P\Sigma^{-1}\mu\} . \quad (3.1)$$

The set $\mathbb{R}^d \times \mathcal{C}_d^+$ is endowed with the scalar product $\langle (a, A), (b, B) \rangle = a^T b + \text{Trace}(A^T B)$. We will use the same notation $\|\cdot\|$ for the norm induced by this scalar product, for the Euclidean norm on \mathbb{R}^d , and for the norm $\|A\| = \text{Tr}(A^T A)^{1/2}$ on $d \times d$ real matrices.

Denote by \mathcal{S}_d the set of $d \times d$ symmetric real matrices; and for $P \in \mathcal{P}$, let $U_P = (I - P)^T(I - P)$. Let $\alpha > 0$ be fixed and define $H : \mathbb{X} \times \Theta \rightarrow \mathbb{R}^d \times \mathcal{S}_d$ by

$$H(x, \theta) = (H_\mu(x, \theta), H_\Sigma(x, \theta)) \quad (3.2)$$

where

$$\begin{aligned} H_\mu(x, \theta) &= x - \mu - \alpha \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} U_P \Sigma^{-1} \mu, \\ H_\Sigma(x, \theta) &= (x - \mu)(x - \mu)^T - \Sigma \\ &\quad + \alpha \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} (\mu \mu^T \Sigma^{-1} U_P + U_P \Sigma^{-1} \mu \mu^T). \end{aligned}$$

Finally, for any $\delta > 0$, set

$$\mathcal{K}_\delta = \{(\mu, \Sigma) \in \Theta : \inf_{P \in \mathcal{P}^*} \|(I - P)\Sigma^{-1}\mu\| \geq \delta\}. \quad (3.3)$$

Let $(\delta_q)_{q \geq 0}$ be a decreasing positive sequence such that $\lim_{q \rightarrow \infty} \delta_q = 0$ and \mathcal{K}_{δ_0} is not empty; choose $\theta_0 = (\mu_0, \Sigma_0) \in \mathcal{K}_{\delta_0}$. Algorithm 2 describes the stable AMOR algorithm in pseudocode.

Algorithm 2.

```

STABLEAMOR( $\pi(\cdot), X_0, T, \theta_0 = (\mu_0, \Sigma_0), c, (\gamma_t)_{t \geq 0}, \alpha, (\mathcal{K}_{\delta_q})_{q \geq 0}$ )
1   $\mathcal{S} \leftarrow \emptyset$ 
2   $\psi \leftarrow 0$   $\triangleright$  Projection counter
3  for  $t \leftarrow 1$  to  $T$ 
4     $\Sigma \leftarrow c\Sigma_{t-1}$   $\triangleright$  scaled adaptive covariance
5     $\tilde{X} \sim \mathcal{N}(\cdot | X_{t-1}, \Sigma)$   $\triangleright$  proposal
6     $\tilde{P} \sim \arg \min_{P \in \mathcal{P}} L_{\theta_{t-1}}(P\tilde{X})$   $\triangleright$  pick an optimal permutation
7     $\tilde{X} \leftarrow \tilde{P}\tilde{X}$   $\triangleright$  permute
8    if  $\frac{\pi(\tilde{X})\sum_P \mathcal{N}(PX_{t-1}|\tilde{X}, \Sigma)}{\pi(X_{t-1})\sum_P \mathcal{N}(P\tilde{X}|X_{t-1}, \Sigma)} > \mathcal{U}[0, 1]$  then
9       $X_t \leftarrow \tilde{X}$   $\triangleright$  accept
10   else
11      $X_t \leftarrow X_{t-1}$   $\triangleright$  reject
12      $\mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$   $\triangleright$  update posterior sample
13      $\mu_t \leftarrow \mu_{t-1} + \gamma_t H_{\mu_{t-1}}(X_t, \theta_{t-1})$ 
14      $\Sigma_t \leftarrow \Sigma_{t-1} + \gamma_t H_{\Sigma_{t-1}}(X_t, \theta_{t-1})$ .
15     if  $(\mu_t, \Sigma_t) \notin \mathcal{K}_{\delta_\psi}$  then
16        $(\mu_t, \Sigma_t) \leftarrow (\mu_0, \Sigma_0)$   $\triangleright$  Project back to  $\mathcal{K}_{\delta_0}$ 
17        $\psi \leftarrow \psi + 1$   $\triangleright$  Increment projection counter
18      $\theta_t \leftarrow (\mu_t, \Sigma_t)$ .
19   return  $\mathcal{S}$ 

```

To prevent that the new value (μ_t, Σ_t) moves into the set $\{\theta \in \Theta : \inf_{P \in \mathcal{P}^*} \|(I - P)\Sigma^{-1}\mu\| = 0\}$, we modify the updates of μ and Σ in Steps 13 and 14 (Steps 12 and 13 in Algorithm 1), and add a projection mechanism in Steps 15 to 17.

We now prove that stable AMOR is an adaptive MCMC algorithm. For any $\theta \in \Theta$, define the transition kernel P_θ on $(\mathbb{X}, \mathcal{X})$ by

$$P_\theta(x, A) = \int_{A \cap V_\theta} \alpha_\theta(x, y) q_\theta(x, y) \, dy + \mathbb{1}_A(x) \int_{V_\theta} (1 - \alpha_\theta(x, z)) q_\theta(x, z) \, dz, \quad (3.4)$$

where V_θ is given by (2.2),

$$\alpha_\theta(x, y) = 1 \wedge \frac{\pi(y) q_\theta(y, x)}{\pi(x) q_\theta(x, y)}, \quad (3.5)$$

and

$$q_\theta(x, y) = \sum_{P \in \mathcal{P}} \mathcal{N}(Py|x, c\Sigma). \quad (3.6)$$

For $\theta \in \Theta$, define also

$$\pi_\theta = |\mathcal{P}| \mathbb{1}_{V_\theta} \pi. \quad (3.7)$$

The following proposition shows that $q_\theta(x, \cdot)$ is a density on V_θ and, the distribution π_θ given by (3.7) is invariant for the transition kernel P_θ . It also establishes that stable AMOR is an adaptive MCMC algorithm: given (X_{t-1}, θ_{t-1}) , X_t is obtained by one iteration of a random walk Metropolis-Hastings algorithm with proposal $q_{\theta_{t-1}}$ and invariant distribution $\pi_{\theta_{t-1}}$.

Proposition 3.1. *Under Assumption 1, the following assertions hold:*

1. For any $\theta \in \Theta$ and $x \in \mathbb{X}$, $\int_{V_\theta} q_\theta(x, y) \, dy = 1$.
2. For any $\theta \in \Theta$, $\pi_\theta P_\theta = \pi_\theta$ and for any $x \in V_\theta$, $P_\theta(x, V_\theta) = 1$.
3. Let $(\theta_t, X_t)_{t \geq 0}$ be given by Algorithm 2. Conditionally on $\sigma(X_0, \theta_0, X_1, \theta_1, \dots, X_{t-1}, \theta_{t-1})$, the distribution of X_t is $P_{\theta_{t-1}}(X_{t-1}, \cdot)$.

Note that the proof of Proposition 3.1 is independent of the update scheme of $(\theta_t)_{t \geq 0}$, which makes the proposition valid for both Algorithms 1 and 2.

3.2. Convergence of stable AMOR

Let

$$\mu_{\pi_\theta} = \int x \pi_\theta(x) \, dx, \quad (3.8)$$

$$\Sigma_{\pi_\theta} = \int (x - \mu_{\pi_\theta})(x - \mu_{\pi_\theta})^T \pi_\theta(x) \, dx, \quad (3.9)$$

be the expectation and covariance matrix of π_θ , respectively. Define the *mean field* $h : \Theta \rightarrow \mathbb{R}^d \times \mathcal{S}_d$ by

$$h(\theta) = (h_\mu(\theta), h_\Sigma(\theta)), \quad (3.10)$$

where

$$\begin{aligned} h_\mu(\theta) &= \mu_{\pi_\theta} - \mu - \alpha \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} U_P \Sigma^{-1} \mu, \\ h_\Sigma(\theta) &= \Sigma_{\pi_\theta} - \Sigma + (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T \\ &\quad + \alpha \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} (\mu \mu^T \Sigma^{-1} U_P + U_P \Sigma^{-1} \mu \mu^T). \end{aligned}$$

The key ingredient for the proof of the convergence of the sequence $(\theta_t)_{t \geq 0}$ is the existence of a Lyapunov function w for the mean field h : we prove in the Appendix (see Lemma 5.6) that the function $w : \Theta \rightarrow \mathbb{R}_+$, defined by

$$w(\theta) = - \int \log \mathcal{N}(x|\theta) \pi_\theta(x) dx + \frac{\alpha}{2} \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^2}, \quad (3.11)$$

is continuously differentiable on Θ and satisfies $\langle \nabla w, h \rangle \leq 0$. In addition, $\langle \nabla w(\theta), h(\theta) \rangle = 0$ iff θ is in the set

$$\mathcal{L} = \{\theta \in \Theta : h(\theta) = 0\} = \{\theta \in \Theta : \nabla w(\theta) = 0\}. \quad (3.12)$$

The convergence of the sequence $(\theta_t)_{t \geq 0}$ is proved by verifying the sufficient conditions for the convergence of the stochastic approximation for Lyapunov stable dynamics given in [1]. The first step is to prove that the sequence is bounded with probability one: we prove that, almost surely, the number of projections ψ is finite so that the projection mechanism (Steps 15 to 17 in Algorithm 2) never occurs after a (random) finite number of iterations. We then prove the convergence of the stable sequence. To achieve that goal, following the same lines as in [1], we make the following assumption.

Assumption 2. *Let \mathcal{L} be given by (3.12). There exists $M_\star > 0$ such that $\mathcal{L} \subset \{\theta : w(\theta) \leq M_\star\}$, and $w(\mathcal{L})$ has an empty interior.*

For $x \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$, define $d(x, A) = \inf_{a \in A} \|x - a\|$. The following result is proved in the Appendix.

Theorem 3.2. *Let $\beta \in (1/2, 1]$ and $\gamma_\star > 0$. Let $(\theta_t)_{t \geq 0}$ be the sequence produced by Algorithm 2 with $\gamma_t \sim \gamma_\star t^{-\beta}$ when $t \rightarrow +\infty$. Under Assumptions 1 and 2,*

1. *Almost surely, there exist $M > 0$ and $t_\star > 0$ such that for any $t \geq t_\star$, $\theta_t \in \{\theta \in \Theta : w(\theta) \leq M\}$. In addition, the number of projections is finite almost surely.*
2. *Almost surely, $(w(\theta_t))_t$ converges to $w^\star \in w(\mathcal{L})$ and $\limsup_t d(\theta_t, \mathcal{L}_{w^\star}) \rightarrow 0$ where $\mathcal{L}_{w^\star} = \{\theta \in \mathcal{L}, w(\theta) = w^\star\}$.*

Theorem 3.2 states the convergence of $(\theta_t)_{t \geq 0}$ to the set \mathcal{L} of the zeros of h ; note that this set neither depends on the initial values (θ_0, X_0) nor on other design parameters. In our experiments, we always observed pointwise convergence. We now state a strong law of large numbers for the samples $(X_t)_{t \geq 0}$, which holds for all paths such that $(\theta_t)_t$ converges to a point $\theta^\star \in \mathcal{L}$.

Theorem 3.3. *Let $\beta \in (1/2, 1]$, $\gamma_\star > 0$, and $\theta^\star \in \mathcal{L}$. Let $(X_t, \theta_t)_{t \geq 0}$ be the sequence generated by Algorithm 2 with $\gamma_t \sim \gamma_\star t^{-\beta}$ when $t \rightarrow +\infty$. Under Assumptions 1 and 2, on the set $\{\lim_t \theta_t = \theta^\star\}$, almost surely,*

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T f(X_t) = \pi_{\theta^\star}(f),$$

for any bounded function f .

Finally, Theorem 3.4 yields the ergodicity of AMOR.

Theorem 3.4. Let $\beta \in (1/2, 1]$, $\gamma_\star > 0$, and $\theta^\star \in \mathcal{L}$. Let $(X_t, \theta_t)_{t \geq 0}$ be the sequence generated by Algorithm 2 with $\gamma_t \sim \gamma_\star t^{-\beta}$ when $t \rightarrow +\infty$. Under Assumptions 1 and 2,

$$\lim_{t \rightarrow \infty} \sup_{\|f\|_\infty \leq 1} \left| \mathbb{E}[f(X_t) \mathbb{1}_{\lim_q \theta_q = \theta^\star}] - \pi_{\theta^\star}(f) \mathbb{P}(\lim_q \theta_q = \theta^\star) \right| = 0.$$

The expression (3.11) of w provides insight into the links between relabeling and vector quantization [13]. The first term is similar to a distortion measure in vector quantization as noted in [5]. It can also be seen as the cross-entropy between π_θ and a Gaussian with parameters θ . The second term in (3.11) is similar to a barrier penalty in continuous optimization [7]. From this perspective, Algorithm 2 can be seen as a constrained optimization procedure that minimizes the cross-entropy. In that sense, if θ^\star denotes a solution to this optimization problem, the *reabeled target* $\pi_{\theta^\star} \propto \mathbb{1}_{V_{\theta^\star}} \pi$ is the restriction of π to one of its symmetric modes V_{θ^\star} that looks as Gaussian as possible among all such restrictions.

Vector quantization algorithms have already been investigated using stochastic approximation tools [21]. However, stability was guaranteed in previous work by making strong assumptions on the trajectories of the process $(\theta_t)_{t \geq 0}$, such as in [21, Theorem 32], see also [21, Results 33 to 37 & Remark 38]. These assumptions ensure that (θ_t) stays asymptotically away from sets where the function used elsewhere as a Lyapunov function is not differentiable. In this paper, we adopt a different strategy by introducing the modifications of the stable AMOR algorithm and adding a barrier term in the definition of our Lyapunov function (3.11) that penalizes these sets. One of the contributions of this paper is to show that this penalization strategy leads to a stable algorithm, without requiring any strong assumption on (θ_t) .

4. Conclusion

We illustrated AMOR, an adaptive Metropolis algorithm with online relabeling that we previously proposed in [5], and proved a strong law of large numbers for a stable version of AMOR. Our algorithm adapts both its proposal and its target on the fly, which makes it a turn-key algorithm. Our results lead to a sound characterization of the target of AMOR that does not depend on the initialization of the algorithm nor on the user. This is the first theoretical analysis of an online relabeling algorithm to our knowledge. The proof further shows how relabeling is related to vector quantization. Unlike previous work on stochastic approximation schemes for vector quantization, we make no strong assumptions on the trajectories of the process considered, rather, we ensure that the appropriate constraint is satisfied by introducing penalization directly into the stochastic approximation framework.

We now examine possible directions for future work. First, following our analysis in Section 3, the question of the control of the convergence of AMOR arises, and proving a central limit theorem would be a natural next step. Second, the online nature of AMOR makes it cheaper than its post-processing counterpart, but it still requires to sweep over all elements of \mathcal{P} at each iteration. This is prohibitive in problems with large $|\mathcal{P}|$, such as additive models with a large number of components. In future work, we will concentrate on algorithmic modifications to reduce this cost, potentially inspired by *probabilistic* relabeling algorithms [17, 31], while conserving our theoretical results. Third, we are interested in extending AMOR to trans-dimensional problems, such as mixtures with an unknown number of components. Reversible jump MCMC (RJMCMC; [14]) also suffers from label-switching and inferential difficulties. We will study algorithms that combine RJMCMC and AMOR.

5. Appendix: proofs

Throughout the proof, let $\Delta_\pi > 0$ be such that

$$x \in \mathbb{X} \Rightarrow \|x\| \leq \Delta_\pi . \quad (5.1)$$

For any function $f : D \rightarrow \mathbb{R}$, we will denote by $\|f\|_\infty = \sup_{x \in D} |f(x)|$.

5.1. Preliminary results

We restate (with a slight adaptation) Lemma 1 of the supplementary material from [5] that we will use extensively.

Lemma 5.1. *Let $\theta \in \Theta$.*

1. *The sets $\{PV_\theta, P \in \mathcal{P}\}$ cover \mathbb{X} , and for any $P, Q \in \mathcal{P}$ such that $P \neq Q$, the Lebesgue measure of $PV_\theta \cap QV_\theta$ is zero.*
2. *Let λ be a measure on $(\mathbb{X}, \mathcal{X})$ with a density w.r.t. the Lebesgue measure. Furthermore, let λ be such that for any $A \in \mathcal{X}$ and $\mathcal{P} \in \mathcal{P}$, $\lambda(PA) = \lambda(A)$. Then $\lambda(V_\theta) = \lambda(\mathbb{X})/|\mathcal{P}|$.*

Proof. (1) Let $\theta \in \Theta$. We first prove that for any $P, Q \in \mathcal{P}$ and $P \neq Q$, the Lebesgue measure of $PV_\theta \cap QV_\theta$ is zero. Observe that $PV_\theta \cap QV_\theta \subseteq \{x : L_\theta(P^T x) = L_\theta(Q^T x)\}$ and $L_\theta(P^T x) = L_\theta(Q^T x)$ iff

$$(x - P\mu)^T P\Sigma^{-1}P^T(x - P\mu) = (x - Q\mu)^T Q\Sigma^{-1}Q^T(x - Q\mu) ,$$

or, equivalently,

$$x^T (P\Sigma^{-1}P^T - Q\Sigma^{-1}Q^T) x - 2\mu^T (\Sigma^{-1}P^T - \Sigma^{-1}Q^T) x = 0 .$$

Then $\{x : L_\theta(P^T x) = L_\theta(Q^T x)\}$ is either a quadratic or a linear hypersurface, and thus of Lebesgue measure zero, except if both $\Sigma^{-1} = R^T \Sigma^{-1} R$ and $\Sigma^{-1} \mu = R \Sigma^{-1} \mu$ with $R = Q^T P$. Since \mathcal{P} is a group, $R \in \mathcal{P}$ and the definition (3.1) of Θ now guarantees that these two conditions never simultaneously hold when $\theta \in \Theta$.

We now prove that $\mathcal{X} \subseteq \bigcup_{P \in \mathcal{P}} PV_\theta$. For any $x \in \mathcal{X}$, there exists $P \in \mathcal{P}$ such that $L_\theta(Px) = \min_{Q \in \mathcal{P}} L_\theta(Qx)$. Then $x \in P^T V_\theta$ and this concludes the proof since \mathcal{P} is a group.

(2) Let $\theta \in \Theta$. Using item (1), it holds that

$$\lambda(\mathbb{X}) = \int_{\mathbb{X}} d\lambda = \sum_{P \in \mathcal{P}} \int_{PV_\theta} d\lambda = \sum_{P \in \mathcal{P}} \int_{V_\theta} d\lambda = |\mathcal{P}| \int_{V_\theta} d\lambda .$$

□

5.2. Differentiating the cross-entropy term in (3.11)

Now, for $\theta \in \Theta$, let

$$\tilde{w}(\theta) = - \int \log \mathcal{N}(x|\theta) \pi_\theta(x) dx . \quad (5.2)$$

Anticipating that we will need to differentiate the function w defined in (3.11), of which \tilde{w} is the first term, we state and prove three lemmas and a proposition that yield the gradient of

\tilde{w} . Lemma 5.2 explicitly reformulates \tilde{w} as a distortion measure in vector quantization [13]. Lemma 5.3 gives the gradient of a distortion measure for generic loss functions L_θ and a generic open set Θ . Its proof is adapted from [13, Lemma 4.10, page 44]. We then show in Lemma 5.4 that Lemma 5.3 applies to the loss function given by (2.1) and the set Θ given by (3.1). Finally, Proposition 5.5 gives an expression of the gradient of \tilde{w} .

Lemma 5.2. *For any $\theta \in \Theta$,*

$$\tilde{w}(\theta) = \frac{1}{2} \ln \det(\Sigma) + \frac{1}{2} \int \min_{P \in \mathcal{P}} L_{(P\mu, P\Sigma P^T)}(x) \pi(x) dx .$$

Proof. Let $\theta \in \Theta$. By definition of \tilde{w} and by Lemma 5.1,

$$\tilde{w}(\theta) = \frac{1}{2} \ln \det(\Sigma) + \frac{|\mathcal{P}|}{2} \int_{V_\theta} L_\theta(x) \pi(x) dx ,$$

where V_θ and L_θ are given respectively by (2.2) and (2.1). Upon noting that π is invariant under the action of \mathcal{P} , we compute

$$|\mathcal{P}| \int_{V_\theta} L_\theta(x) \pi(x) dx = \sum_{P \in \mathcal{P}} \int_{V_\theta} L_\theta(x) \pi(x) dx = \sum_{P \in \mathcal{P}} \int_{PV_\theta} L_\theta(P^T x) \pi(x) dx .$$

In addition, by the definition (2.2) of V_θ ,

$$PV_\theta = \{x \in \mathcal{X} : L_\theta(P^T x) = \min_{Q \in \mathcal{P}} L_\theta(Qx)\} .$$

Then by Lemma 5.1,

$$|\mathcal{P}| \int_{V_\theta} L_\theta(x) \pi(x) dx = \sum_{P \in \mathcal{P}} \int_{PV_\theta} \min_{Q \in \mathcal{P}} L_\theta(Qx) \pi(x) dx = \int \min_{Q \in \mathcal{P}} L_\theta(Qx) \pi(x) dx .$$

Finally, by the definition (2.1) of L_θ , $L_\theta(Qx) = L_{(Q^T \mu, Q^T \Sigma Q)}(x)$, and this concludes the proof. \square

Lemma 5.3. *Let Θ be an open subset of \mathbb{R}^ℓ , r be a positive integer and $\mathcal{O} \subseteq \Theta^r$ be an open set. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a measurable set and π be a probability density w.r.t. the Lebesgue measure on \mathcal{X} . Let $\{L_\theta, \theta \in \Theta\}$ be a family of loss functions $L_\theta : \mathcal{X} \rightarrow \mathbb{R}$, satisfying*

A. *For π -almost every x , $\theta \mapsto L_\theta(x)$ is C^1 on Θ and for any $\theta \in \Theta$, there exists $h_0 > 0$ such that*

$$\int \sup_{\|h\| \leq h_0} \frac{1}{\|h\|} |h^T \nabla_\theta L_\theta(x)| \pi(x) dx < \infty .$$

B. *For any $\theta \in \Theta$, there exists $h_0 > 0$ such that*

$$\int \sup_{\|h\| \leq h_0} \frac{|L_{\theta+h}(x) - L_\theta(x)|}{\|h\|} \pi(x) dx < \infty .$$

C. *For any $\theta = (\theta_1, \dots, \theta_r) \in \mathcal{O}$, the sets*

$$V_{\theta_i} = \{x \in \mathcal{X} : L_{\theta_i}(x) \leq \min_j L_{\theta_j}(x)\}$$

are measurable, cover \mathcal{X} and for any $i \neq j$, the Lebesgue measure of $V_{\theta_i} \cap V_{\theta_j}$ is zero.

For $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r) \in \mathcal{O}$ define the function $\varphi : \Theta^r \rightarrow \mathbb{R}$ by

$$\varphi(\boldsymbol{\theta}) = \int \min_{1 \leq i \leq r} L_{\theta_i}(x) \pi(x) dx .$$

Then φ is differentiable on \mathcal{O} and for $1 \leq i \leq r$,

$$\nabla_{\theta_i} \varphi(\boldsymbol{\theta}) = \int_{V_{\theta_i}} \nabla_{\theta_i} L_{\theta_i}(x) \pi(x) dx .$$

Proof. Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_r) \in \mathcal{O}$. Set

$$d(x, \boldsymbol{\theta}) = \min_{1 \leq i \leq r} L_{\theta_i}(x) .$$

By definition of the function φ

$$\varphi(\boldsymbol{\theta} + \mathbf{h}) - \varphi(\boldsymbol{\theta}) = \int (d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta})) \pi(x) dx . \quad (5.3)$$

We now prove that

$$\lim_{\|h\| \rightarrow 0} \|h\|^{-1} \left(\varphi(\boldsymbol{\theta} + \mathbf{h}) - \varphi(\boldsymbol{\theta}) - \sum_{i=1}^r \int_{V_{\theta_i}} \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \pi(x) dx \right) = 0$$

by applying the dominated convergence theorem. First, by Assumption C,

$$\begin{aligned} \varphi(\boldsymbol{\theta} + \mathbf{h}) - \varphi(\boldsymbol{\theta}) &= \sum_{i=1}^r \int_{V_{\theta_i}} \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \pi(x) dx \\ &= \sum_{i=1}^r \int_{V_{\theta_i}} (d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle) \pi(x) dx . \end{aligned}$$

Now set

$$V_{\theta_i}^\circ = \{x \in \mathcal{X} : L_{\theta_i}(x) < \min_{j \neq i} L_{\theta_j}(x)\}$$

and note that $V_{\theta_i} \setminus V_{\theta_i}^\circ$ has measure zero under Assumption C. Then

$$\begin{aligned} \varphi(\boldsymbol{\theta} + \mathbf{h}) - \varphi(\boldsymbol{\theta}) &= \sum_{i=1}^r \int_{V_{\theta_i}} \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \pi(x) dx \\ &= \sum_{i=1}^r \int_{V_{\theta_i}^\circ} (d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle) \pi(x) dx . \end{aligned}$$

Let $x \in V_{\theta_i}^\circ$; under Assumption A, $\theta \mapsto L_\theta(x)$ is continuous on Θ and there exists ε_x such that

$$\|h\| \leq \varepsilon_x \Rightarrow d(x, \boldsymbol{\theta} + \mathbf{h}) = L_{\theta_i + h_i}(x) .$$

Then, by Assumption A,

$$\begin{aligned} d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle &= L_{\theta_i + h_i}(x) - L_{\theta_i}(x) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \\ &= C(\theta_i, x, h_i) \end{aligned}$$

with $\|h_i\|^{-1}C(\theta_i, x, h_i) \rightarrow 0$ when $\|h_i\| \rightarrow 0$. Hence, we proved that for any $i \leq r$ and any $x \in V_{\theta_i}^\circ$,

$$\lim_{\|h\| \rightarrow 0} \|h\|^{-1} (d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) - \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle) = 0.$$

We now prove that there exists h_0 such that

$$\int \sup_{\|h\| \leq h_0} \|h\|^{-1} |d(x, \boldsymbol{\theta} + \mathbf{h}) - d(x, \boldsymbol{\theta}) - \sum_{i=1}^r \langle \nabla_{\theta_i} L_{\theta_i}(x), h_i \rangle \mathbb{1}_{V_{\theta_i}}(x)| \pi(x) dx < +\infty. \quad (5.4)$$

First remark that for all $z, \mathbf{a} = (a_1, \dots, a_r), \mathbf{b} = (b_1, \dots, b_r)$,

$$|d(z, \mathbf{a} + \mathbf{b}) - d(z, \mathbf{a})| \leq \max_{1 \leq i \leq r} |L_{a_i+b_i}(z) - L_{a_i}(z)|. \quad (5.5)$$

Indeed, assume without loss of generality that $d(z, \mathbf{a}) \leq d(z, \mathbf{a} + \mathbf{b})$ and let i be such that $d(z, \mathbf{a}) = L_{a_i}(z)$, then by definition of the distance d , $d(z, \mathbf{a} + \mathbf{b}) \leq L_{a_i+b_i}(z)$, which proves Eq. (5.5). Now, the proof of (5.4) is a consequence of Assumptions A and B and the inequality

$$\max_{1 \leq i \leq r} |L_{a_i+b_i}(z) - L_{a_i}(z)| \leq \sum_{i=1}^r |L_{a_i+b_i}(z) - L_{a_i}(z)|.$$

□

Lemma 5.4. *Under Assumption 1, the quadratic loss function given by (2.1), the set Θ given by (3.1), and the open set*

$$\mathcal{O} = \{(P\mu, P\Sigma P^T) : P \in \mathcal{P}, (\mu, \Sigma) \in \Theta\}$$

satisfy the assumptions of Lemma 5.3.

Proof. When taking derivatives with respect to a matrix, we shall use the “vec” notation during computations. For a $d \times d$ matrix A , its vectorized form $\text{vec}(A)$ is a d^2 vector such that $\text{vec}(A)$ stacks the columns of A on top of one another. In general, we refer to [8] for matrix algebra notions.

We check the conditions of Lemma 5.3. Denote by r the cardinality of \mathcal{P} and set $\mathcal{P} = (I_d, P_2, \dots, P_r)$, where I_d is the $d \times d$ identity matrix. We set

$$\mathcal{O} = \{(\theta_1, \dots, \theta_r) \in \Theta^r : \theta_i = (P_i\mu, P_i\Sigma P_i^T), \forall i \geq 1\}.$$

Note that for $\boldsymbol{\theta} \in \mathcal{O}$, $L_{\theta_i}(x) = L_{\theta_1}(P_i^T x)$ and $V_{\theta_i} = P_i V_{\theta_1}$. Now, we have

$$(\mu, \Sigma) \mapsto (x - \mu)^T \Sigma^{-1} (x - \mu) = \frac{1}{\det \Sigma} (x - \mu)^T \text{Adjugate}(\Sigma) (x - \mu)$$

so that $\theta \mapsto L_\theta(x)$ is a rational function in the coefficients of μ and Σ whose denominator $\det \Sigma > 0$. In addition,

$$\sup_{\|h\| \leq h_0} \frac{1}{\|h\|} |h^T \nabla_\theta L_\theta(x)| \leq \|\nabla_\theta L_\theta(x)\| \leq \|\nabla_\mu L_\theta(x)\| + \|\nabla_\Sigma L_\theta(x)\|.$$

The RHS is at most quadratic in x (for fixed θ). By Assumption 1, the RHS is π -integrable. This proves Assumption A of Lemma 5.3.

We now prove Assumption B of Lemma 5.3. Let $\theta \in \Theta$ and set $\Delta\theta = (\Delta\mu, \Delta\Sigma)$. By standard algebra, we have

$$(\Sigma + \Delta\Sigma)^{-1} = \Sigma^{-1} - \Sigma^{-1} \Delta\Sigma \Sigma^{-1} + o(\|\Delta\Sigma\|)$$

for any matrix $\Delta\Sigma$ such that $\Sigma + \Delta\Sigma$ is invertible. Therefore,

$$L_{\theta+\Delta\theta}(x) - L_\theta(x) = -2(\Delta\mu)^T \Sigma^{-1}(x - \mu) - (x - \mu)^T \Sigma^{-1} \Delta\Sigma \Sigma^{-1}(x - \mu) + \Xi(x, \theta, \Delta\theta),$$

for some function $\Xi(x, \theta, \Delta\theta)$ such that

$$|\Xi(x, \theta, \Delta\theta)| \leq C(\theta) \|x\|^2 \|\Delta\theta\|^2$$

and some constant $C(\theta)$ (depending upon θ but independent of x and $\Delta\theta$). The proof is concluded since, by Assumption 1, $\int \|x\|^2 \pi(x) dx < +\infty$.

Finally, the sets V_{θ_i} are measurable for any $\theta_1, \dots, \theta_r \in \Theta$ since $(x, \theta) \mapsto L_\theta(x)$ is continuous on $\mathcal{X} \times \Theta$. The proof of Assumption C of Lemma 5.3 is then concluded by application of Lemma 5.1. \square

We are now ready to state the final result of this preliminary section, and give the expression of the gradient of \tilde{w} defined in (5.2).

Proposition 5.5. *Under Assumption 1, the function \tilde{w} defined in (5.2) is continuously differentiable on Θ and for any $\theta \in \Theta$,*

$$\begin{aligned} \nabla_\mu \tilde{w}(\theta) &= -\Sigma^{-1}(\mu_{\pi_\theta} - \mu), \\ \nabla_\Sigma \tilde{w}(\theta) &= -\frac{1}{2} \Sigma^{-1} (\Sigma_{\pi_\theta} - \Sigma + (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T) \Sigma^{-1}. \end{aligned}$$

Proof. Let r denote the cardinality of \mathcal{P} and set $\mathcal{P} = (I_d, P_2, \dots, P_r)$. Let $\theta \in \Theta$. By Lemma 5.2, we have

$$\tilde{w}(\theta) = \frac{1}{2} \ln \det(\Sigma) + \frac{1}{2} \int \min_{1 \leq i \leq r} L_{\theta_i}(x) \pi(x) dx,$$

where $\theta_i = (P_i \mu, P_i \Sigma^{-1} P_i^T)$.

We first consider the derivative w.r.t. μ . We have

$$\nabla_\mu \tilde{w}(\theta) = \frac{1}{2} \nabla_\mu \int \min_{1 \leq i \leq r} L_{\theta_i}(x) \pi(x) dx.$$

By Lemmas 5.3 and 5.4 and the chain rule, we have

$$\begin{aligned} \nabla_\mu \tilde{w}(\theta) &= \frac{1}{2} \sum_{i=1}^r P_i^T \int_{A_i} \nabla_{\mu_i} [(x - \mu_i) P_i \Sigma^{-1} P_i^T (x - \mu_i)]_{\mu_i = P_i \mu} \pi(x) dx \\ &= -\Sigma^{-1} \sum_{i=1}^r \int_{A_i} (P_i^T x - \mu) \pi(x) dx, \end{aligned}$$

where

$$A_i = \{x : L_{\theta_i}(x) \leq \min_j L_{\theta_j}(x)\} = P_i V_\theta,$$

with V_θ defined in (2.2). Hence, by Lemma 5.1, and since π is invariant under the action of \mathcal{P} , we have

$$\begin{aligned}\nabla_\mu \tilde{w}(\theta) &= -\Sigma^{-1} \sum_{i=1}^r \int_{V_\theta} (x - \mu) \pi(x) dx \\ &= -\Sigma^{-1} \int (x - \mu) [r \pi(x) \mathbb{1}_{V_\theta}(x)] dx \\ &= -\Sigma^{-1} (\mu_{\pi_\theta} - \mu) ,\end{aligned}$$

where we used the definition (3.8) of μ_{π_θ} .

We now consider the derivative w.r.t. Σ , that we will derive in a similar manner. We refer to [8] for matrix algebra notions such as Kronecker products. First remark that, by standard algebra and since Σ is symmetric,

$$\nabla_{\text{vec}(\Sigma)} \ln \det \Sigma = \text{vec}(\Sigma^{-1}) .$$

Then recall that

$$\nabla_{\text{vec}(\Sigma)} (x - \mu) \Sigma^{-1} (x - \mu) = -\Sigma^{-1} (x - \mu) \otimes \Sigma^{-1} (x - \mu) .$$

Now let, for A a matrix, $A^{\otimes 2} = A \otimes A$. Using Lemmas 5.3 and 5.4 along with the chain rule, we compute

$$\begin{aligned}\nabla_{\text{vec}(\Sigma)} \tilde{w}(\theta) - \frac{1}{2} \text{vec}(\Sigma^{-1}) &= \frac{1}{2} \sum_{i=1}^r (P_i^{\otimes 2})^T \int_{P_i V_\theta} \nabla_{\text{vec}(\Sigma_i)} [(x - P_i \mu)^T \Sigma_i^{-1} (x - P_i \mu)]_{\Sigma_i = P_i \Sigma P_i^T} \pi(x) dx \\ &= -\frac{1}{2} \sum_{i=1}^r (P_i^T)^{\otimes 2} \int_{P_i V_\theta} [P_i \Sigma^{-1} P_i^T (x - P_i \mu)]^{\otimes 2} \pi(x) dx \\ &= -\frac{1}{2} \sum_{i=1}^r \int_{P_i V_\theta} [\Sigma^{-1} (P_i^T x - \mu)]^{\otimes 2} \pi(x) dx \\ &= -\frac{1}{2} (\Sigma^{-1})^{\otimes 2} \sum_{i=1}^r \int_{P_i V_\theta} [P_i^T x - \mu]^{\otimes 2} \pi(x) dx\end{aligned}$$

where we used the identities $(A \otimes B)^T = A^T \otimes B^T$ and $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$. A

change of variables now leads to

$$\begin{aligned}
& \nabla_{\text{vec}(\Sigma)} \tilde{w}(\theta) - \frac{1}{2} \text{vec}(\Sigma^{-1}) \\
&= -\frac{1}{2} (\Sigma^{-1})^{\otimes 2} \sum_{i=1}^r \int_{V_\theta} (x - \mu)^{\otimes 2} \pi(x) dx \\
&= -\frac{1}{2} (\Sigma^{-1})^{\otimes 2} \int (x - \mu_{\pi_\theta} + \mu_{\pi_\theta} - \mu)^{\otimes 2} [r \pi(x) \mathbb{1}_{V_\theta}(x)] dx \\
&= -\frac{1}{2} (\Sigma^{-1})^{\otimes 2} \left(\int (x - \mu_{\pi_\theta}) \otimes (x - \mu_{\pi_\theta}) \pi_\theta(x) dx + (\mu_{\pi_\theta} - \mu) \otimes (\mu_{\pi_\theta} - \mu) \right) \\
&= -\frac{1}{2} (\Sigma^{-1} \otimes \Sigma^{-1}) \text{vec}(\Sigma_{\pi_\theta} + (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T),
\end{aligned}$$

where we used the distributivity of the Kronecker product, Lemma 5.1, and the definitions (3.8) and (3.9) of μ_{π_θ} and Σ_{π_θ} . Finally, the identity $\text{vec}(AXB) = (B^T \otimes A) \text{vec}(X)$ allows us to write

$$\nabla_{\text{vec}(\Sigma)} \tilde{w}(\theta) = -\frac{1}{2} \text{vec}(\Sigma^{-1} [\Sigma_{\pi_\theta} - \Sigma + (\mu_{\pi_\theta} - \mu)(\mu_{\pi_\theta} - \mu)^T] \Sigma^{-1}).$$

□

5.3. The Lyapunov function

Lemma 5.6 establishes the existence of a Lyapunov function for the mean field h given by (3.10).

Lemma 5.6. *Under Assumption 1, the mean field h is continuous on Θ , the function w defined by (3.11) is C^1 on Θ and*

1. $\nabla_\mu w(\theta) = -\Sigma^{-1} h_\mu(\theta)$ and $\nabla_\Sigma w(\theta) = -\frac{1}{2} \Sigma^{-1} h_\Sigma(\theta) \Sigma^{-1}$.
2. $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$ on Θ and $\langle \nabla w(\theta), h(\theta) \rangle = 0$ iff $\theta \in \mathcal{L}$.
3. For any $M > 0$, the level set

$$\mathcal{W}_M = \{\theta \in \Theta : w(\theta) \leq M\} \tag{5.6}$$

is a compact subset of Θ , and there exist $\delta_1, \delta_2 > 0$ such that

$$\inf_{\theta \in \mathcal{W}_M} \inf_{P \in \mathcal{P}^*} \|(I - P) \Sigma^{-1} \mu\| \geq \delta_1 \text{ and} \tag{5.7a}$$

$$\inf_{\theta \in \mathcal{W}_M} \lambda_{\min}(\Sigma) \geq \delta_2, \tag{5.7b}$$

where $\lambda_{\min}(\Sigma)$ denotes the minimal eigenvalue of the real symmetric matrix Σ .

Remark 5.7. As a consequence of Lemma 5.6, observe that for any $M > 0$, there exists $\delta > 0$ such that $\mathcal{W}_M \subseteq \mathcal{K}_\delta$, where \mathcal{K}_δ is defined in (3.3).

Proof. (Continuity of h) Since $(I - P) \Sigma^{-1} \mu \neq 0$ on Θ for any $P \in \mathcal{P}^*$, it suffices to show that $\theta \mapsto \mu_{\pi_\theta}$ and $\theta \mapsto \Sigma_{\pi_\theta}$ are continuous. Since, by Lemma 5.1, the boundary of V_θ is of Lebesgue measure zero, the continuity of $\theta \mapsto \mu_{\pi_\theta}$ follows from Lebesgue's dominated convergence theorem if, for any $x \in \mathbb{X} \setminus \partial V_\theta$, $\theta \mapsto x \mathbb{1}_{V_\theta}(x)$ is continuous. To see this, note that if x is in the interior of

V_θ , then there exists a neighborhood \mathcal{V} of θ such that for any $\theta' \in \mathcal{V}$, $x \in V_{\theta'}$, and if $x \in \mathbb{X} \setminus V_\theta$, which is an open subset of \mathbb{X} , then there exists a neighborhood \mathcal{V} of θ such that for any $\theta' \in \mathcal{V}$, $x \in \mathbb{X} \setminus V_{\theta'}$.

The case of $\theta \mapsto \Sigma_\theta$ is similar and omitted.

(w is C^1 on Θ) It is shown in [5, Proposition 3 of the supplementary material] that the first term in the RHS of (3.11) is continuously differentiable on Θ . Since $\|(I - P)\Sigma^{-1}\mu\| \neq 0$ for any $P \in \mathcal{P}^*$ and $(\mu, \Sigma) \in \Theta$, the second term in the RHS of (3.11) is continuously differentiable on Θ . By [5, Proposition 3 of the supplementary material], it holds for any $\theta = (\mu, \Sigma) \in \Theta$ that

$$\begin{aligned} \nabla_\mu w(\theta) &= -\Sigma^{-1}(\mu_{\pi_\theta} - \mu) + \alpha \sum_P \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} \Sigma^{-1} U_P \Sigma^{-1} \mu = -\Sigma^{-1} h_\mu(\theta) \\ \nabla_\Sigma w(\theta) &= -\frac{1}{2} \Sigma^{-1} (\Sigma_{\pi_\theta} - \Sigma + (\mu - \mu_{\pi_\theta})(\mu - \mu_{\pi_\theta})^T) \Sigma^{-1} \dots \\ &\quad - \frac{\alpha}{2} \sum_P \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^4} \Sigma^{-1} (\mu \mu^T \Sigma^{-1} U_P) \Sigma^{-1} + U_P \Sigma^{-1} \mu \mu^T \\ &= -\frac{1}{2} \Sigma^{-1} h_\Sigma(\theta) \Sigma^{-1}. \end{aligned}$$

Hence, upon noting that $h_\Sigma(\theta)$ and Σ^{-1} are symmetric,

$$\begin{aligned} \langle \nabla w(\theta), h(\theta) \rangle &= -h_\mu(\theta)^T \Sigma^{-1} h_\mu(\theta) - \frac{1}{2} \text{Trace}(\Sigma^{-1} h_\Sigma(\theta) \Sigma^{-1} h_\Sigma(\theta)) \\ &= -h_\mu(\theta)^T \Sigma^{-1} h_\mu(\theta) - \frac{1}{2} \text{Trace}(\Sigma^{-1/2} h_\Sigma(\theta) \Sigma^{-1} h_\Sigma(\theta) \Sigma^{-1/2}). \end{aligned}$$

The first term of the RHS is negative since $\Sigma \in \mathcal{C}_d^+$ and the second term is negative since $(A, B) \mapsto \text{Trace}(A^T B)$ is a scalar product. Therefore $\langle \nabla w(\theta), h(\theta) \rangle \leq 0$ with equality iff $\theta \in \mathcal{L}$.

(\mathcal{W}_M is compact) We prove (5.7a). By the definition (3.11) of w , for any $\theta \in \mathcal{W}_M$, we have

$$-\int \log \mathcal{N}(x|\theta) \pi_\theta(x) dx + \frac{\alpha}{2} \sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^2} \leq M.$$

In particular, the first term in the LHS is a cross-entropy, and it is thus non-negative (alternatively, see [5, Proposition 1 of the supplementary material]). Consequently, for any $\theta \in \mathcal{W}_M$, we have

$$\sum_{P \in \mathcal{P}^*} \frac{1}{\|(I - P)\Sigma^{-1}\mu\|^2} \leq \frac{2M}{\alpha}.$$

This yields $\|(I - P)\Sigma^{-1}\mu\|^2 \geq \frac{\alpha}{2M}$ for any $P \in \mathcal{P}^*$, thus concluding the proof of (5.7a).

We now prove (5.7b). Let $\theta = (\mu, \Sigma) \in \mathcal{W}_M$. Denote by $(\lambda_i(\Sigma))_{i \leq d}$ the eigenvalues of Σ . Since Σ is symmetric, there exist $d \times d$ matrices Q_θ, Λ_θ such that $\Sigma = Q_\theta \Lambda_\theta Q_\theta^T$, Q_θ is orthogonal, and $\Lambda_\theta = \text{diag}(\lambda_i(\Sigma))$. Then

$$\begin{aligned} 2M &\geq 2w(\theta) \geq -2 \int \log \mathcal{N}(x|\theta) \pi_\theta(x) dx \\ &= d \log(2\pi) + \log \det \Sigma + (\mu_{\pi_\theta} - \mu)^T \Sigma^{-1} (\mu_{\pi_\theta} - \mu) + \text{Trace}(\Sigma^{-1} \Sigma_{\pi_\theta}) \\ &\geq \sum_{i=1}^d \log \lambda_i(\theta) + 0 + \text{Trace}(\Sigma^{-1} \Sigma_{\pi_\theta}). \end{aligned} \tag{5.8}$$

Set $b_i(\theta) = (Q_\theta^T \Sigma_{\pi_\theta} Q_\theta)_{ii}$. Then

$$\text{Trace}(\Sigma^{-1} \Sigma_{\pi_\theta}) = \text{Trace}(Q_\theta \Lambda_\theta^{-1} Q_\theta^T \Sigma_{\pi_\theta}) = \text{Trace}(Q_\theta^T \Sigma_{\pi_\theta} Q_\theta \Lambda_\theta^{-1}) = \sum_{i=1}^d \frac{b_i(\theta)}{\lambda_i(\theta)}. \quad (5.9)$$

Therefore, for any $\theta \in \mathcal{W}_M$,

$$\sum_{i=1}^d \log \lambda_i(\theta) + \frac{b_i(\theta)}{\lambda_i(\theta)} \leq 2M. \quad (5.10)$$

We now prove that for any i , $\inf_{\mathcal{W}_M} b_i > 0$. This property, combined with (5.10), will conclude the proof of (5.7b). Let $\epsilon > 0$ be such that $2^d \epsilon \|\pi\|_\infty \Delta_\pi^{d-1} < |\mathcal{P}|$, and for $v \in \{x \in \mathbb{R}^d : \|x\| = 1\}$, let

$$B_\epsilon^v(\theta) = \{x \in \text{Supp}(\pi) \cap V_\theta : |\langle x - \mu_{\pi_\theta}, v \rangle| \leq \epsilon\}. \quad (5.11)$$

Note that by Assumption 1,

$$\pi(B_\epsilon^v(\theta)) \leq \|\pi\|_\infty \text{Leb}(B_\epsilon^v(\theta)) \leq 2^d \epsilon \|\pi\|_\infty \Delta_\pi^{d-1}.$$

Then, by definition of ϵ ,

$$\pi(V_\theta \setminus B_\epsilon^v(\theta)) \geq |\mathcal{P}| - 2^d \epsilon \|\pi\|_\infty \Delta_\pi^{d-1} > 0. \quad (5.12)$$

Now, if (e_i) denotes the canonical basis of \mathbb{R}^d , then

$$\begin{aligned} b_i(\theta) &= |\mathcal{P}| e_i^T Q_\theta^T \left(\int_{V_\theta} (x - \mu_{\pi_\theta})(x - \mu_{\pi_\theta})^T \pi(x) dx \right) Q_\theta e_i \\ &= |\mathcal{P}| \int_{V_\theta} (Q_\theta e_i)^T (x - \mu_{\pi_\theta})(x - \mu_{\pi_\theta})^T Q_\theta e_i \pi(x) dx \\ &= |\mathcal{P}| \int_{V_\theta} \langle x - \mu_{\pi_\theta}, Q_\theta e_i \rangle^2 \pi(x) dx \\ &\geq |\mathcal{P}| \int_{V_\theta \setminus B_\epsilon^{Q_\theta e_i}(\theta)} \langle x - \mu_{\pi_\theta}, Q_\theta e_i \rangle^2 \pi(x) dx \\ &\geq \epsilon^2 |\mathcal{P}| \pi(V_\theta \setminus B_\epsilon^{Q_\theta e_i}(\theta)), \end{aligned} \quad (5.13)$$

where the last inequality follows from the definition (5.11) of $B_\epsilon^{Q_\theta e_i}(\theta)$. Thus, by (5.12), $b_i(\theta)$ is bounded away from zero on \mathcal{W}_M .

As w is continuous on Θ , $\{\theta \in \Theta, w(\theta) \leq M\}$ is closed. From (5.7b), (5.8) and Assumption 1, $\mu \mapsto (\mu_{\pi_\theta} - \mu)^T \Sigma^{-1} (\mu_{\pi_\theta} - \mu)$ is bounded on \mathcal{W}_M . In addition, (5.8), (5.9) and (5.13) imply that $\Sigma \mapsto \log \det \Sigma$ is bounded on \mathcal{W}_M . These properties combined with (5.7b) imply that \mathcal{W}_M is bounded. Hence \mathcal{W}_M is compact. \square

5.4. Proof of Proposition 3.1

(1) By the definition (3.1) of Θ and Lemma 5.1, $\forall \theta \in \Theta, x \in \mathbb{X}$, it holds that

$$\int_{V_\theta} q_\theta(x, y) dy = \sum_{P \in \mathcal{P}} \int_{V_\theta} \mathcal{N}(Py|x, c\Sigma) dy = 1.$$

(2) Let $(X_t)_{t \geq 0}$ and $(\theta_t)_{t \geq 0}$ be the random processes defined by Algorithm 2. We prove that for any measurable positive function f ,

$$\mathbb{E}[f(X_t)|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] = \int f(x_t) P_{\theta_{t-1}}(X_{t-1}, x_t) dx_t, w.p.1.$$

Let f be measurable and positive. Let (\tilde{P}, \tilde{X}) be the r.v. defined by Steps 5 and 6. Let U be a uniform r.v. independent of $\sigma(X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}, \tilde{P}, \tilde{X})$. By construction, it holds that

$$\begin{aligned} \mathbb{E}[f(X_t)|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] &= \mathbb{E}[f(\tilde{P}\tilde{X})\mathbb{1}_{U \leq \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})}|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] \\ &\quad + \mathbb{E}[f(X_{t-1})\mathbb{1}_{U > \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})}|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}], \end{aligned} \quad (5.14)$$

where $\alpha_\theta(x, y)$ is given by (3.5). Since U is independent of the past and from \tilde{P} and \tilde{X} , we have

$$\begin{aligned} \mathbb{E}[f(\tilde{P}\tilde{X})\mathbb{1}_{U \leq \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})}|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] \\ = \mathbb{E}\left[f(\tilde{P}\tilde{X})\left(1 - \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})\right)|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}\right], \end{aligned} \quad (5.15)$$

and

$$\begin{aligned} \mathbb{E}[f(X_{t-1})\mathbb{1}_{U > \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})}|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] \\ = f(X_{t-1}) \mathbb{E}\left[\left(1 - \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})\right)|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}\right]. \end{aligned} \quad (5.16)$$

Now note that the projection mechanism (Steps 15 to 17 of Algorithm 2) guarantees that $\theta_{t-1} \in \Theta$ with probability 1. By Lemma 5.1, $\theta \in \Theta$ implies $\mathbb{X} = \cup_P(PV_\theta)$ and

$$\forall P, Q \in \mathcal{P} \text{ such that } P \neq Q, \text{ Leb}(PV_\theta \cap QV_\theta) = 0.$$

Thus, for any measurable and bounded function $\varphi : \mathbb{X} \times \Theta \rightarrow \mathbb{R}$, we have

$$\int_{\mathbb{X}} \varphi(x, \theta) dx = \sum_{Q \in \mathcal{P}} \int_{QV_\theta \cap (\cup_{R \neq Q} RV_\theta)^c} \varphi(x, \theta) dx.$$

Applying this decomposition to (5.15) yields

$$\begin{aligned} \mathbb{E}[f(\tilde{P}\tilde{X})\mathbb{1}_{U \leq \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})}|X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] \\ = \sum_{P \in \mathcal{P}} \int h(Px) \frac{1}{N(x, \theta_{t-1})} \mathbb{1}_{V_{\theta_{t-1}}}(Px) \mathcal{N}(x|X_{t-1}, c\Sigma_{t-1}) dx \\ = \sum_{P, Q \in \mathcal{P}} \int_{QV_{\theta_{t-1}} \cap (\cup_{R \neq Q} RV_{\theta_{t-1}})^c} h(Px) \frac{1}{N(x, \theta_{t-1})} \mathbb{1}_{V_{\theta_{t-1}}}(Px) \mathcal{N}(x|X_{t-1}, c\Sigma_{t-1}) dx, \end{aligned}$$

where $N(x, \theta) = |\{Q \in \mathcal{P} | Qx \in V_\theta\}|$. Using Lemma 5.1 again,

$$\theta \in \Theta, x \notin \cup_{P \neq Q}(PV_\theta \cap QV_\theta) \Rightarrow N(x, \theta) = 1,$$

and thus

$$\begin{aligned}
& \mathbb{E}[f(\tilde{P}\tilde{X})\mathbb{1}_{U \leq \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X})} | X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1}] \\
&= \sum_{P, Q \in \mathcal{P}} \int_{Q V_{\theta_{t-1}} \cap (\cup_{R \neq Q} R V_{\theta_{t-1}})^c} h(Px) \mathbb{1}_{V_{\theta_{t-1}}}(Px) \mathcal{N}(x | X_{t-1}, c\Sigma_{t-1}) dx \\
&= \sum_{P \in \mathcal{P}} \int h(Px) \mathbb{1}_{V_{\theta_{t-1}}}(Px) \mathcal{N}(x | X_{t-1}, c\Sigma_{t-1}) dx \\
&= \sum_{P \in \mathcal{P}} \int h(y) \mathbb{1}_{V_{\theta_{t-1}}}(y) \mathcal{N}(P^{-1}y | X_{t-1}, c\Sigma_{t-1}) dy \\
&= \int_{V_{\theta_{t-1}}} h(y) q_{\theta_{t-1}}(X_{t-1}, y) dy,
\end{aligned}$$

where in the last step we used the fact that \mathcal{P} is a group. Similarly,

$$\begin{aligned}
& \mathbb{E} \left[\left(1 - \alpha_{\theta_{t-1}}(X_{t-1}, \tilde{P}\tilde{X}) \right) | X_0, \theta_0, \dots, X_{t-1}, \theta_{t-1} \right] \\
&= \int_{V_{\theta_{t-1}}} \left((1 - \alpha_{\theta_{t-1}}(X_{t-1}, y)) \right) q_{\theta_{t-1}}(X_{t-1}, y) dy;
\end{aligned}$$

and this concludes the proof.

(3) Let $\theta \in \Theta$. Eqn. (3.4) implies that if $x \in V_\theta$, then $P(x, V_\theta) = 1$. To prove that $\pi_\theta P_\theta = \pi_\theta$, it is sufficient to check the detailed balance condition, which states that

$$\forall A, B \subset \mathbb{X} \text{ measurable}, \int_A \pi_\theta(x) P_\theta(x, B) dx = \int_B \pi_\theta(y) P_\theta(y, A) dy.$$

We consider the two summands in the definition (3.4) separately. First, it holds that

$$\begin{aligned}
\pi_\theta(x) \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) &= |P| (\pi(x) q_\theta(x, y) \wedge \pi(y) q_\theta(y, x)) \mathbb{1}_{V_\theta}(x) \mathbb{1}_{V_\theta}(y) \\
&= \pi_\theta(y) \alpha_\theta(y, x) q_\theta(y, x) \mathbb{1}_{V_\theta}(x),
\end{aligned}$$

so

$$\int_A \pi_\theta(x) \left(\int_{B \cap V_\theta} \alpha_\theta(x, y) q_\theta(x, y) dy \right) dx = \int_B \pi_\theta(y) \left(\int_{A \cap V_\theta} \alpha_\theta(y, x) q_\theta(y, x) dx \right) dy.$$

Secondly,

$$\begin{aligned}
& \int_A \pi_\theta(x) \mathbb{1}_B(x) \int_{V_\theta} (1 - \alpha_\theta(x, z)) q_\theta(x, z) dz dx \\
&= \int_{A \cap B} \pi_\theta(x) \int_{V_\theta} (1 - \alpha_\theta(x, z)) q_\theta(x, z) dz dx \\
&= \int_B \pi_\theta(y) \mathbb{1}_A(y) \int_{V_\theta} (1 - \alpha_\theta(y, z)) q_\theta(y, z) dz.
\end{aligned}$$

This concludes the proof of the detailed balance condition.

5.5. Regularity in θ of the Poisson solution

Lemma 5.8.

1. For any $M > 0$, there exists $\rho \in (0, 1)$ such that for any $x \in \mathbb{X}$ and any $\theta \in \mathcal{W}_M$, $\|P_\theta^n(x, \cdot) - \pi_\theta\|_{\text{TV}} \leq 2(1 - \rho)^n$.
2. Under Assumption 1, for any $\theta \in \Theta$, there exists a solution \hat{H}_θ of the Poisson equation $g - P_\theta g = H(\cdot, \theta) - \pi_\theta H(\cdot, \theta)$. Furthermore, for any $M > 0$,

$$\sup_{\theta \in \mathcal{W}_M} \sup_{x \in \mathbb{X}} |\hat{H}_\theta(x)| < \infty. \quad (5.17)$$

Proof. (of Item 1) It is sufficient to prove that there exists $\rho \in (0, 1)$ such that for any $x \in \mathbb{X}$ and $\theta \in \mathcal{W}_M$, $P_\theta(x, \cdot) \geq \rho \pi_\theta$ (see e.g. [20, Theorem 16.2.4]). By (3.4), for any $x \in \mathbb{X}$ and $A \in \mathcal{X}$, $P_\theta(x, A) \geq \int_{A \cap V_\theta} \alpha_\theta(x, y) q_\theta(x, y) dy$. By Lemma 5.6, there exists $a > 0$ such that for any $(\mu, \Sigma) \in \mathcal{W}_M$, any $m, z \in \mathbb{X}$, and any $P \in \mathcal{P}$, we have $\mathcal{N}(Pz|m, \Sigma) \geq a$. Thus, for any $\theta \in \mathcal{W}_M$ and $y \in V_\theta$, it holds that

$$\alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) \geq a |\mathcal{P}| \left(1 \wedge \frac{\pi(y)}{\pi(x)} \right) \mathbb{1}_{V_\theta}(y) \geq \frac{a}{\|\pi\|_\infty} \pi_\theta(y). \quad (5.18)$$

Thus, we have $P_\theta(x, \cdot) \geq \rho \pi_\theta$ for any $x \in \mathbb{X}$ and $\theta \in \mathcal{W}_M$ with $\rho = a/\|\pi\|_\infty$.

(Proof of Item 2)

$$\begin{aligned} \left| \sum_n P_\theta^n(H(x, \theta) - \pi_\theta(H(\cdot, \theta))) \right| &\leq \sup_{\theta \in \mathcal{W}_M} \|H(\cdot, \theta)\|_\infty \sum_n \|P_\theta^n(x, \cdot) - \pi_\theta\|_{\text{TV}} \\ &\leq 2 \sup_{\theta \in \mathcal{W}_M} \|H(\cdot, \theta)\|_\infty \rho^{-1}. \end{aligned} \quad (5.19)$$

Since the sup is finite by Lemma 5.6, the series $\sum P_\theta^n(H(x, \theta) - \pi_\theta(H(\cdot, \theta)))$ converges. Finally, note that

$$\hat{H}_\theta(x) = \sum_n P_\theta^n(H(x, \theta) - \pi_\theta(H(\cdot, \theta)))$$

is a solution of the Poisson equation, and that $\sup_{\theta \in \mathcal{W}_M, x \in \mathbb{X}} |\hat{H}_\theta(x)| < \infty$. \square

Lemma 5.9. Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \Theta$, it holds that

$$\text{Leb}(V_\theta \setminus V_{\theta'}) \leq C \|\theta - \theta'\|^{1-2\kappa}, \quad (5.20)$$

where $\text{Leb}(A)$ denotes the Lebesgue measure of the set A .

Proof. We prove that there exist $\bar{C}, \bar{h} > 0$, such that for any $\theta \in \mathcal{W}_M$ and any $\theta' \in \Theta$ such that $\|\theta - \theta'\| \leq \bar{h}$, $\text{Leb}(V_\theta \setminus V_{\theta'}) \leq \bar{C} \|\theta - \theta'\|^{1-2\kappa}$. Note that since $V_\theta \subset \mathbb{X}$ and since \mathbb{X} is bounded, there exists $\check{C} > 0$ such that $\text{Leb}(V_\theta \setminus V_{\theta'}) \leq \check{C}$. Therefore, (5.20) holds with $C = \bar{C} \vee \check{C}/\bar{h}^{1-2\kappa}$.

By Lemma 5.6, w is uniformly continuous on \mathcal{W}_{M+1} , and there exists $h_0 > 0$ small enough for which

$$[\theta \in \mathcal{W}_M, \theta' \in \Theta, \|\theta - \theta'\| < h_0] \Rightarrow \forall u \in [0, 1], \theta + u(\theta' - \theta) \in \mathcal{W}_{M+1}. \quad (5.21)$$

Let $\bar{h} \leq h_0$. Let $\theta = (\mu, \Sigma) \in \mathcal{W}_M$ and $\theta' \neq \theta$ such that $\|\theta - \theta'\| \leq \bar{h}$.

By definition of the set V_ϑ , for any $x \in V_\theta \setminus V_{\theta'}$, there exists $P \in \mathcal{P}^*$ such that $L_{\theta'}(x) - L_{\theta'}(P^T x) > 0$ and $L_\theta(x) - L_\theta(P^T x) \leq 0$. Since $\vartheta \mapsto L_\vartheta(x) - L_\vartheta(P^T x)$ is continuous on \mathcal{W}_{M+1} , there exists $u \in [0, 1]$ depending on x, θ, θ' , and P such that $L_{\theta+u(\theta'-\theta)}(x) - L_{\theta+u(\theta'-\theta)}(P^T x) = 0$. Therefore

$$V_\theta \setminus V_{\theta'} \subset \bigcup_{P \in \mathcal{P}^*} \mathcal{V}_P,$$

where

$$\mathcal{V}_P = \bigcup_{u \in [0, 1]} \mathcal{Z}(L_{\theta+u(\theta'-\theta)}(\cdot) - L_{\theta+u(\theta'-\theta)}(P^T \cdot)) \cap \mathbb{X}; \quad (5.22)$$

and $\mathcal{Z}(f)$ denotes the zeros of the function f . The proof proceeds by showing that for any $P \in \mathcal{P}^*$, \mathcal{V}_P is included in a measurable set with measure $O(\|\theta - \theta'\|^{1-2\kappa})$.

Let $P \in \mathcal{P}^*$. Let $B(0, \Delta_\pi) = \{y \in \mathbb{R}^d : \|y\| \leq \Delta_\pi\}$, where Δ_π is defined by 5.1. For any $x \in B(0, \Delta_\pi)$, define

$$\begin{aligned} l_\theta(x) &= 2\mu^T \Sigma^{-1}(I - P^T)x, \\ q_\theta(x) &= x^T(\Sigma^{-1} - P\Sigma^{-1}P^T)x, \\ \mathcal{B}_{\theta, \theta'} &= \{x \in B(0, \Delta_\pi) : |l_\theta(x)| \leq \|\theta - \theta'\|^\kappa\}. \end{aligned}$$

Denote by \mathbb{S} the unit sphere $\{x \in \mathbb{R}^d : \|x\| = 1\}$. Let $u \in [0, 1]$ and $tv \in \mathcal{Z}(L_{\theta+u(\theta'-\theta)}(\cdot) - L_{\theta+u(\theta'-\theta)}(P^T \cdot)) \cap \mathbb{X}$ where $t \in [0, \Delta_\pi]$ and $v \in \mathbb{S}$. Upon noting that for any $\vartheta \in \mathcal{W}_{M+1}$,

$$L_\vartheta(tv) - L_\vartheta(tP^T v) = t(q_\vartheta(v)t - l_\vartheta(v)) \quad , \quad (5.23)$$

we consider several cases:

- (i) $tv \in \mathcal{B}_{\theta, \theta'}$.
- (ii) $tv \notin \mathcal{B}_{\theta, \theta'}$ and $q_{\theta+u(\theta'-\theta)}(v) = 0$. Then, by (5.23), $l_{\theta+u(\theta'-\theta)}(tv) = 0$ which implies that $tv \in \mathcal{B}_{\theta, \theta'}$. This yields a contradiction.
- (iii) $tv \notin \mathcal{B}_{\theta, \theta'}$ and $q_{\theta+u(\theta'-\theta)}(v) \neq 0$. Then $t \neq 0$ and, by (5.23),

$$t = \frac{l_{\theta+u(\theta'-\theta)}(v)}{q_{\theta+u(\theta'-\theta)}(v)}. \quad (5.24)$$

Since we assumed $t \in [0, \Delta_\pi]$, this ratio is positive. In order to characterize the point tv , additional notations are required. First, note that by Lemma 5.6, there exists $C_1 > 0$ such that for any $\tilde{\theta} = (\tilde{\mu}, \tilde{\Sigma}) \in \mathcal{W}_{M+1}$,

$$\|\tilde{\theta} - \theta\| \leq h_0 \Rightarrow \|\tilde{\Sigma}^{-1} - \Sigma^{-1}\| \leq C_1 \|\tilde{\Sigma} - \Sigma\|.$$

Thus, there exists $C_2 > 0$ such that for any $\tilde{\theta} \in \mathcal{W}_{M+1}$, $\|\tilde{\theta} - \theta\| \leq h_0$, and for any $x \in B(0, \Delta_\pi)$,

$$\begin{aligned} |l_{\tilde{\theta}}(x) - l_\theta(x)| &= 2 \left| \mu^T [\tilde{\Sigma}^{-1} - \Sigma^{-1}](I - P^T)x + (\tilde{\mu} - \mu)^T \tilde{\Sigma}^{-1}(I - P^T)x \right| \\ &\leq C_2 \|\tilde{\theta} - \theta\|. \end{aligned} \quad (5.25)$$

Note that since $x, \mu \in B(0, \Delta_\pi)$, C_2 does not depend on x and θ . Similarly, there exists $C_3 > 0$ such that for $x \in B(0, \Delta_\pi)$ and $\tilde{\theta} \in \mathcal{W}_{M+1}$ satisfying $\|\tilde{\theta} - \theta\| \leq h_0$,

$$|q_{\tilde{\theta}}(x) - q_\theta(x)| \leq C_3 \|\tilde{\theta} - \theta\|. \quad (5.26)$$

We can assume without loss of generality that \bar{h} is small enough so that

$$\|\theta - \theta'\| \leq \bar{h} \Rightarrow \|\theta - \theta'\|^\kappa - (C_2 + 2C_3\Delta_\pi) \|\theta - \theta'\| \geq \frac{1}{2} \|\theta - \theta'\|^\kappa. \quad (5.27)$$

We now distinguish three subcases.

a) $v \in \mathcal{B}_{\theta, \theta'}$.

b) $v \notin \mathcal{B}_{\theta, \theta'}$ and $q_\theta(v) \neq 0$. Since $t \in [0, \Delta_\pi]$, (5.24) implies that $|q_{\theta+u(\theta'-\theta)}(v)| \geq |l_{\theta+u(\theta'-\theta)}(v)|/\Delta_\pi$. Since $v \notin \mathcal{B}_{\theta, \theta'}$, $|l_\theta(v)| \geq \|\theta - \theta'\|^\kappa$ and by using (5.25),

$$|l_{\theta+u(\theta'-\theta)}(v)| \geq |l_\theta(v)| - |l_{\theta+u(\theta'-\theta)} - l_\theta(v)| \geq \|\theta - \theta'\|^\kappa - C_2 \|\theta - \theta'\|.$$

Hence, it holds that $|q_{\theta+u(\theta'-\theta)}(v)| \geq (\|\theta - \theta'\|^\kappa - C_2 \|\theta - \theta'\|)/\Delta_\pi$, and, by (5.26), we have $|q_\theta(v)| \geq |q_{\theta+u(\theta'-\theta)}(v)| - C_3 \|\theta - \theta'\|$. These inequalities together with (5.25) and (5.27) lead to

$$\left| t - \frac{l_\theta(v)}{q_\theta(v)} \right| = \left| \frac{l_{\theta+u(\theta'-\theta)}(v)}{q_{\theta+u(\theta'-\theta)}(v)} - \frac{l_\theta(v)}{q_\theta(v)} \right| \leq C_4 \|\theta - \theta'\|^{1-2\kappa},$$

for some $C_4 > 0$.

c) $v \notin \mathcal{B}_{\theta, \theta'}$ and $q_\theta(v) = 0$. Then by (5.25) and (5.26),

$$t \geq \frac{\|\theta - \theta'\|^\kappa - C_2 \|\theta - \theta'\|}{C_3 \|\theta - \theta'\|} \geq 2\Delta_\pi,$$

which is in contradiction with the assumption that $t \leq \Delta_\pi$.

As a conclusion, we have just proved that \mathcal{V}_P is included in the union of three sets defined by $\mathcal{B}_{\theta, \theta'}$ (case i), by $\{tv : t \in [0, \Delta_\pi], v \in \mathbb{S} \cap \mathcal{B}_{\theta, \theta'}\}$ (case iia), and by

$$\left\{ tv : v \in \mathbb{S}, v \notin \mathcal{B}_{\theta, \theta'}, q_\theta(v) \neq 0, 0 \leq t \leq \Delta_\pi, \left| t - \frac{l_\theta(v)}{q_\theta(v)} \right| \leq C_4 \|\theta - \theta'\|^{1-2\kappa} \right\}$$

(case iic). This concludes the first step.

The second step consists in computing an upper bound for the Lebesgue measure of each of these three sets. For simplifying the presentation, we detail the case $d = 2$ and use polar coordinates (ρ, ϕ) ; the argument remains valid when $d > 2$ using generalized spherical coordinates. Define $t_\theta(\phi) = l_\theta(e^{i\phi})/q_\theta(e^{i\phi})$. Rephrasing the conclusion of the first step, we have $\mathcal{V}_P \subset \bigcup_{\ell=1}^3 \mathcal{V}_P^{(\ell)}$ with

$$\mathcal{V}_P^{(1)} = \mathcal{B}_{\theta, \theta'},$$

$$\mathcal{V}_P^{(2)} = \{(\rho, \phi) / \rho \in [0, \Delta_\pi], e^{i\phi} \in \mathcal{B}_{\theta, \theta'}\},$$

$$\mathcal{V}_P^{(3)} = \{(\rho, \phi) / e^{i\phi} \notin \mathcal{B}_{\theta, \theta'}, q_\theta(e^{i\phi}) \neq 0, 0 \leq \rho \leq \Delta_\pi, |\rho - t_\theta(\phi)| \leq C_4 \|\theta - \theta'\|^{1-2\kappa}\}.$$

These sets are Borel sets. By definition of \mathcal{W}_M , l_θ is not identically zero and thus

$$\text{Leb}(\mathcal{V}_P^{(1)}) = \text{Leb}(\mathcal{B}_{\theta, \theta'}) \leq 2\Delta_\pi \frac{\|\theta - \theta'\|^{1-2\kappa}}{\|2\mu^t \Sigma^{-1}(I - PT)\|} \leq C_5 \|\theta - \theta'\|^{1-2\kappa}$$

for some $C_5 > 0$ as a consequence of Lemma 5.6. For $\mathcal{V}_P^{(2)}$, note that it is upper bounded by the reunion of the two circular sectors in bold lines in Figure 7. This area is easily bounded by the area of the outer rectangle, which is proportional to $\|\theta - \theta'\|^{1-2\kappa}$. Finally,

$$\text{Leb}(\mathcal{V}_P^{(3)}) = \int_0^{2\pi} \left[\frac{\rho^2}{2} \right]_{0 \vee (t_\theta(\phi) - C_4 \|\theta - \theta'\|^{1-2\kappa})}^{\Delta_\pi \wedge (t_\theta(\phi) + C_4 \|\theta - \theta'\|^{1-2\kappa})} \mathbb{1}_{q_\theta(e^{i\phi}) \neq 0} d\phi.$$

We can assume without loss of generality that \bar{h} is small enough so that $2C_4 \bar{h}^{1-2\kappa} < \Delta_\pi$. Therefore, we can partition $[0, 2\pi] = \mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$, where

$$\begin{aligned} \mathcal{A} &= \{ \phi \in [0, 2\pi] / t_\theta(\phi) - C_4 \|\theta - \theta'\|^{1-2\kappa} \geq 0 \text{ and } t_\theta(\phi) + C_4 \|\theta - \theta'\|^{1-2\kappa} \leq \Delta_\pi \}, \\ \mathcal{B} &= \{ \phi \in [0, 2\pi] / t_\theta(\phi) - C_4 \|\theta - \theta'\|^{1-2\kappa} \geq 0 \text{ and } t_\theta(\phi) + C_4 \|\theta - \theta'\|^{1-2\kappa} \geq \Delta_\pi \}, \\ \mathcal{C} &= \{ \phi \in [0, 2\pi] / t_\theta(\phi) - C_4 \|\theta - \theta'\|^{1-2\kappa} \leq 0 \text{ and } 0 \leq t_\theta(\phi) + C_4 \|\theta - \theta'\|^{1-2\kappa} \leq \Delta_\pi \}. \end{aligned}$$

This yields

$$\begin{aligned} \text{Leb}(\mathcal{V}_P^{(3)}) &\leq 2C_4 \int_{\mathcal{A}} t_\theta(\phi) \|\theta - \theta'\|^{1-2\kappa} d\phi + \frac{1}{2} \int_{\mathcal{B}} \left(\Delta_\pi^2 - (t_\theta(\phi) - C_4 \|\theta - \theta'\|^{1-2\kappa})^2 \right) d\phi \\ &\quad + \frac{1}{2} \int_{\mathcal{C}} (t_\theta(\phi) + C_4 \|\theta - \theta'\|^{1-2\kappa})^2 d\phi \end{aligned} \quad (5.28)$$

$$\leq C_6 \|\theta - \theta'\|^{1-2\kappa}, \quad (5.29)$$

for some $C_6 > 0$, since on \mathcal{A} , $0 \leq t_\theta(\phi) \leq \Delta_\pi$, on \mathcal{B} , $(t_\theta(\phi) - C_4 \|\theta - \theta'\|^{1-2\kappa})^2 \geq (\Delta_\pi - 2C_4 \|\theta - \theta'\|^{1-2\kappa})^2$, and on \mathcal{C} , $|t_\theta(\phi)| \leq C_4 \|\theta - \theta'\|^{1-2\kappa}$.

This concludes the proof. \square

Lemma 5.10. (Regularity in θ of the invariant distribution π_θ)

Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \Theta$,

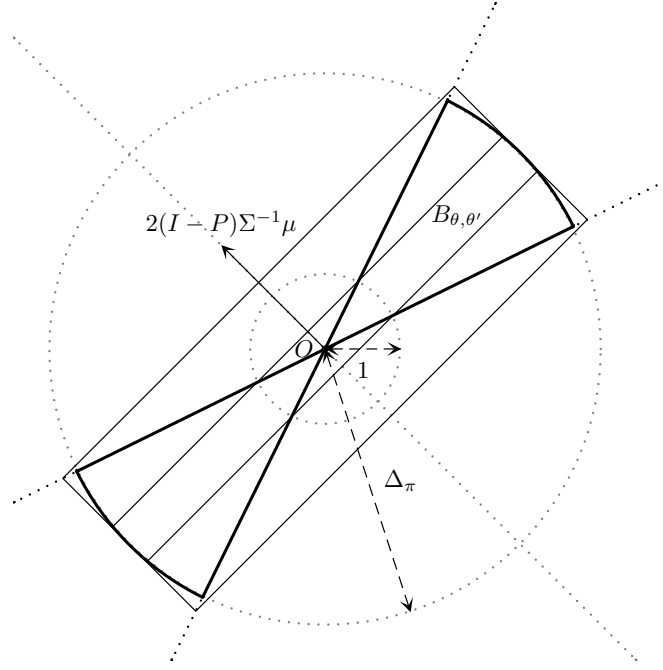
$$\|\pi_\theta - \pi_{\theta'}\|_{\text{TV}} \leq C \|\theta - \theta'\|^{1-2\kappa}.$$

Proof. By definition of the total variation,

$$\begin{aligned} \|\pi_\theta - \pi_{\theta'}\|_{\text{TV}} &= \sup_{\|f\|_\infty \leq 1} \left| \int f(x) \pi_\theta(x) dx - \int f(x) \pi_{\theta'}(x) dx \right| \\ &= |\mathcal{P}| \sup_{\|f\|_\infty \leq 1} \left| \int_{V_\theta \setminus V_{\theta'}} f(x) \pi(x) dx - \int_{V_{\theta'} \setminus V_\theta} f(x) \pi(x) dx \right| \\ &\leq |\mathcal{P}| (\pi(V_\theta \setminus V_{\theta'}) + \pi(V_{\theta'} \setminus V_\theta)). \end{aligned}$$

Since

$$V_{\theta'} \setminus V_\theta = V_\theta \setminus (V_\theta \cap V_{\theta'}), \quad V_\theta \setminus V_{\theta'} = V_\theta \setminus (V_\theta \cap V_{\theta'}),$$

FIG 7. Bounding the measure of the set $\mathcal{V}_P^{(2)}$.

it holds that

$$\pi(V_{\theta'} \setminus V_{\theta}) = \frac{1}{|\mathcal{P}|} - \pi(V_{\theta} \cap V_{\theta'}) = \pi(V_{\theta} \setminus V_{\theta'}),$$

where we used Lemma 5.1. Then, by Assumption 1 and Lemma 5.9, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \Theta$,

$$\|\pi_{\theta} - \pi_{\theta'}\|_{\text{TV}} \leq 2\|\pi\|_{\infty} \text{Leb}(V_{\theta} \setminus V_{\theta'}) \leq C\|\theta - \theta'\|^{1-2\kappa}.$$

□

Lemma 5.11. (Regularity in θ of the kernels P_{θ})

Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \mathcal{W}_{M+1}$,

$$\|P_{\theta}(x, \cdot) - P_{\theta'}(x, \cdot)\|_{\text{TV}} \leq C\|\theta - \theta'\|^{1-2\kappa}.$$

Proof. From the definition of the transition kernel P_θ , we have

$$\begin{aligned}
|P_\theta f(x) - P_{\theta'} f(x)| &\leq \left| \int f(y) \left(\alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right) dy \right| \\
&\quad + |f(x)| \left| \int \left(\alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) - \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) \right) dy \right| \\
&\leq 2\|f\|_\infty \int \left| \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right| dy \\
&= 2\|f\|_\infty \sum_{i=1}^4 \Delta_{\theta, \theta'}^i(x), \tag{5.30}
\end{aligned}$$

where

$$\begin{aligned}
\Delta_{\theta, \theta'}^1(x) &= \int_{\mathcal{A}_\theta(x) \cap \mathcal{A}_{\theta'}(x)} \left| \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right| dy, \\
\Delta_{\theta, \theta'}^2(x) &= \int_{\mathcal{R}_\theta(x) \cap \mathcal{R}_{\theta'}(x)} \left| \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right| dy, \\
\Delta_{\theta, \theta'}^3(x) &= \int_{\mathcal{A}_\theta(x) \cap \mathcal{R}_{\theta'}(x)} \left| \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right| dy, \\
\Delta_{\theta, \theta'}^4(x) &= \int_{\mathcal{R}_\theta(x) \cap \mathcal{A}_{\theta'}(x)} \left| \alpha_\theta(x, y) q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \alpha_{\theta'}(x, y) q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right| dy,
\end{aligned}$$

and

$$\mathcal{A}_\theta(x) = \{y : \alpha_\theta(x, y) = 1\}, \quad \mathcal{R}_\theta(x) = \{y : \alpha_\theta(x, y) < 1\}.$$

We now upper bound each term.

$$\begin{aligned}
\Delta_{\theta, \theta'}^1(x) &= \int_{\mathcal{A}_\theta(x) \cap \mathcal{A}_{\theta'}(x)} \left| \sum_{Q \in \mathcal{P}} \left(\mathbb{1}_{V_\theta}(y) \mathcal{N}(Qy|x, \Sigma) - \mathbb{1}_{V_{\theta'}}(y) \mathcal{N}(Qy|x, \Sigma') \right) \right| dy \\
&\leq \int \left| \mathbb{1}_{V_\theta}(y) - \mathbb{1}_{V_{\theta'}}(y) \right| \sum_{Q \in \mathcal{P}} \mathcal{N}(Qy|x, \Sigma) + \\
&\quad \mathbb{1}_{V_{\theta'}}(y) \sum_{Q \in \mathcal{P}} \left| \mathcal{N}(Qy|x, \Sigma) - \mathcal{N}(Qy|x, \Sigma') \right| dy. \tag{5.31}
\end{aligned}$$

By Lemma 5.6, there exist $a, b > 0$ such that for any $\theta \in \mathcal{W}_{M+1}$, $m, z \in \mathbb{X}$, and $Q \in \mathcal{P}$, we have

$$a \leq \mathcal{N}(Qz|m, c\Sigma) \leq b, \tag{5.32}$$

so that the first term in the RHS of (5.31) is bounded by

$$\begin{aligned}
\int \left| \mathbb{1}_{V_\theta}(y) - \mathbb{1}_{V_{\theta'}}(y) \right| \sum_{Q \in \mathcal{P}} \mathcal{N}(Qy|x, \Sigma) dy &\leq |\mathcal{P}|b \int \left| \mathbb{1}_{V_\theta}(y) - \mathbb{1}_{V_{\theta'}}(y) \right| dy \\
&= |\mathcal{P}|b \int (\mathbb{1}_{V_\theta \setminus V_{\theta'}}(y) + \mathbb{1}_{V_{\theta'} \setminus V_\theta}(y)) dy \\
&\leq C\|\theta - \theta'\|^{1-2\kappa},
\end{aligned}$$

where we used Lemma 5.9. Let us now consider the second term of the right-hand side of (5.31). Using the uniform continuity of w on \mathcal{W}_{M+1} (see Lemma 5.6), there exists \bar{h} small enough such that

$$\theta \in \mathcal{W}_M, \|h\| < \bar{h} \Rightarrow \theta + h \in \mathcal{W}_{M+1}. \quad (5.33)$$

For any $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$ such that $\|\theta - \theta'\| \geq \bar{h}$, there exists C_1 such that

$$\sum_{Q \in \mathcal{P}} |\mathcal{N}(Qy|x, \Sigma) - \mathcal{N}(Qy|x, \Sigma')| dy \leq C_1 \|\theta - \theta'\|^{1-2\kappa}.$$

Assume now that $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$ and $\|\theta - \theta'\| < \bar{h}$. Denote by

$$\Sigma_t = (1-t)\Sigma + t\Sigma'. \quad (5.34)$$

By (5.33) and (5.7b), Σ_t^{-1} exists and $\sup_{t \leq 1, \theta \in \mathcal{W}_M, \theta' \in \mathcal{W}_{M+1}} \|\Sigma_t^{-1}\| < \infty$. We can then write

$$\begin{aligned} |\mathcal{N}(Qy|x, \Sigma) - \mathcal{N}(Qy|x, \Sigma')| &= \int_0^1 \mathcal{N}(Qy|x, \Sigma_t) \left| \frac{d}{dt} \log \mathcal{N}(Qy|x, \Sigma_t) \right| dt \\ &\leq b \int_0^1 \left| \frac{d}{dt} \log \mathcal{N}(Qy|x, \Sigma_t) \right| dt. \end{aligned} \quad (5.35)$$

In addition, by Assumption 1, there exists C_2 such that

$$\left| \frac{d}{dt} \log \mathcal{N}(Qy|x, \Sigma_t) \right| = \left| (x - Qy)^T \Sigma_t^{-1} (\Sigma' - \Sigma) \Sigma_t^{-1} (x - Qy) \right| \leq C_2 \|\theta - \theta'\|. \quad (5.36)$$

We thus have proved that

$$[\theta \in \mathcal{W}_M, \theta' \in \mathcal{W}_{M+1}, \|\theta - \theta'\| < \bar{h}] \implies |\mathcal{N}(Qy|x, \Sigma) - \mathcal{N}(Qy|x, \Sigma')| \leq C \|\theta - \theta'\|.$$

Therefore, it is established that $\|\Delta_{\theta, \theta'}^1\|_\infty \leq C \|\theta - \theta'\|^{1-2\kappa}$.

Let us consider the second term $\Delta_{\theta, \theta'}^2(x)$ in the RHS of (5.30). Note first that if $x \in \mathbb{X}$ and $y \in \mathcal{R}_\theta(x) \cap \mathcal{R}_{\theta'}(x)$, then by (5.32), $\pi(y)/\pi(x) \leq b/a$, so

$$\begin{aligned} \Delta_{\theta, \theta'}^2(x) &= \int_{\mathcal{R}_\theta(x) \cap \mathcal{R}_{\theta'}(x)} \frac{\pi(y)}{\pi(x)} \left| \sum_{Q \in \mathcal{P}} \left(\mathbb{1}_{V_\theta}(y) \mathcal{N}(Qx|y, \Sigma) - \mathbb{1}_{V_{\theta'}}(y) \mathcal{N}(Qx|y, \Sigma') \right) \right| dy \\ &\leq \frac{b}{a} \int_{\mathcal{R}_\theta(x) \cap \mathcal{R}_{\theta'}(x)} \left| \sum_{Q \in \mathcal{P}} \left(\mathbb{1}_{V_\theta}(y) \mathcal{N}(Qx|y, \Sigma) - \mathbb{1}_{V_{\theta'}}(y) \mathcal{N}(Qx|y, \Sigma') \right) \right| dy. \end{aligned}$$

Therefore, repeating the above discussion for the bound of $\Delta_{\theta, \theta'}^1(x)$, it is established that $\|\Delta_{\theta, \theta'}^2\|_\infty \leq C \|\theta - \theta'\|^{1-2\kappa}$.

To deal with $\Delta_{\theta, \theta'}^3(x)$, first observe that there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$, and $x, y \in \mathbb{X}$, we have

$$\left| \frac{q_\theta(y, x)}{q_\theta(x, y)} - \frac{q_{\theta'}(y, x)}{q_{\theta'}(x, y)} \right| \leq C \|\theta - \theta'\|, \quad (5.37)$$

because of (3.6), (5.32), and the above discussion for the upper bound of $\Delta_{\theta, \theta'}^1(x)$. Now let $y \in \mathcal{A}_\theta(x) \cap \mathcal{R}_{\theta'}(x)$, then we have

$$\frac{\pi(y)q_{\theta'}(y, x)}{\pi(x)q_{\theta'}(x, y)} \leq 1 \leq \frac{\pi(y)q_\theta(y, x)}{\pi(x)q_\theta(x, y)},$$

which, combined with (5.37), yields

$$1 - C \frac{\pi(y)}{\pi(x)} \|\theta - \theta'\| \leq \frac{\pi(y)q_{\theta'}(y, x)}{\pi(x)q_{\theta'}(x, y)} \leq 1.$$

Thus,

$$\begin{aligned} \Delta_{\theta, \theta'}^3(x) &= \int_{\mathcal{A}_\theta(x) \cap \mathcal{R}_{\theta'}(x)} \left| q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - \frac{\pi(y)q_{\theta'}(y, x)}{\pi(x)q_{\theta'}(x, y)} q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) \right| dy \\ &\leq \int \left(|q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y)| \vee \dots \right. \\ &\quad \left. |q_\theta(x, y) \mathbb{1}_{V_\theta}(y) - q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y) + C \frac{\pi(y)}{\pi(x)} \|\theta - \theta'\| q_{\theta'}(x, y) \mathbb{1}_{V_{\theta'}}(y)| \right) dy. \end{aligned}$$

Therefore, it is established that $\|\Delta_{\theta, \theta'}^3\|_\infty \leq C \|\theta - \theta'\|^{1-2\kappa}$.

The upper bound of $\Delta_{\theta, \theta'}^4(x)$ is similar and thus its proof is omitted. \square

Lemma 5.12. (Regularity in θ of the solution of the Poisson equation)

Let $M > 0$ and $\kappa \in (0, 1/2)$. Under Assumption 1, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \mathcal{W}_{M+1}$,

$$\|P_\theta \hat{H}_\theta - P_{\theta'} \hat{H}_{\theta'}\|_\infty \leq C \|\theta - \theta'\|^{1-2\kappa}.$$

Proof. We recall the following result, proved in [12, Lemma 5.5, page 24]: there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$, $\theta' \in \mathcal{W}_{M+1}$, and $x \in \mathbb{X}$,

$$\begin{aligned} \|P_\theta \hat{H}_\theta - P_{\theta'} \hat{H}_{\theta'}\|_\infty &\leq C \|H(\cdot, \theta) - H(\cdot, \theta')\|_\infty + C \sup_{\theta \in \mathcal{W}_M} \|H(\cdot, \theta)\|_\infty \{ \|\pi_\theta - \pi_{\theta'}\|_{\text{TV}} \\ &\quad + \sup_{x \in \mathbb{X}} \|P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\|_{\text{TV}} \}. \end{aligned} \quad (5.38)$$

Here $\sup_{\theta \in \mathcal{W}_M} \|H(\cdot, \theta)\|_\infty$ is finite by Lemma 5.6. Now, by Lemma 5.6 again, there exists $C > 0$ such that for any $\theta \in \mathcal{W}_M$ and $\theta' \in \mathcal{W}_{M+1}$,

$$\|H(\cdot, \theta) - H(\cdot, \theta')\|_\infty \leq C \|\theta - \theta'\|.$$

The upper bounds for the two last terms in the RHS of (5.38) result from Lemmas 5.10 and 5.11, respectively. \square

5.6. Proof of Theorem 3.2

We start by proving two lemmas.

Lemma 5.13. *Let $(\gamma_t)_{t>0}$ be a sequence such that $\sum_t \gamma_t^2 < \infty$, $\sum_t |\gamma_{t+1} - \gamma_t| < \infty$, and $\sum_t \gamma_t^{2(1-\kappa)} < \infty$ for some $\kappa \in (0, 1/2)$. Denote by ψ_t the value of the projection counter at the end of iteration t , in Algorithm 2. Let $(\theta_t, X_t)_{t \geq 0}$ be the sequence generated by Algorithm 2. Under Assumptions 1 and 2, for any $M > 0$,*

$$\lim_{L \rightarrow +\infty} \sup_{\ell \geq 1} \left\| \left(\prod_{k=L}^{L+\ell} \mathbb{1}_{\theta_k \in \mathcal{W}_M} \mathbb{1}_{\psi_{k+1}=\psi_k} \right) \sum_{k=L}^{L+\ell} \gamma_{k+1} (H(X_{k+1}, \theta_k) - h(\theta_k)) \right\| = 0 \quad w.p.1, \quad (5.39)$$

where H , h , w , and \mathcal{W}_M are given by (3.2), (3.10), (3.11), and (5.6), respectively.

Proof. Let $M > 0$. By uniform continuity of w on \mathcal{W}_{M+1} , let $L(M)$ be large enough so that

$$L \geq L(M), \theta \in \mathcal{W}_M \implies \forall x \in \mathbb{X}, \theta + \gamma_{L+1} H(x, \theta) \in \mathcal{W}_{M+1}. \quad (5.40)$$

Let $L \geq L(M)$ and let

$$\mathbb{I}_{L,\ell} = \prod_{k=L}^{L+\ell} \mathbb{1}_{\theta_k \in \mathcal{W}_M} \mathbb{1}_{\psi_{k+1}=\psi_k}.$$

For any $\theta \in \mathcal{W}_M$, Lemma 5.8 implies that there exists a function \hat{H}_θ such that

$$\hat{H}_\theta - P_\theta \hat{H}_\theta = H(\cdot, \theta) - \pi_\theta(H(\cdot, \theta)) \quad \text{and} \quad \sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_M} \|\hat{H}_\theta(x)\| < \infty.$$

Therefore, for $\ell + L \geq i \geq L \geq 0$, we have

$$\mathbb{I}_{L,\ell} (H(X_{i+1}, \theta_i) - h(\theta_i)) = \mathbb{I}_{L,\ell} (M_{i+1} + R_{i+1}^{(1)} + R_{i+1}^{(2)}),$$

where

$$M_{i+1} = \left(\hat{H}_{\theta_i}(X_{i+1}) - P_{\theta_i} \hat{H}_{\theta_i}(X_i) \right), \quad (5.41)$$

$$R_{i+1}^{(1)} = P_{\theta_i} \hat{H}_{\theta_i}(X_i) - P_{\theta_{i+1}} \hat{H}_{\theta_{i+1}}(X_{i+1}), \quad (5.42)$$

$$R_{i+1}^{(2)} = P_{\theta_{i+1}} \hat{H}_{\theta_{i+1}}(X_{i+1}) - P_{\theta_i} \hat{H}_{\theta_i}(X_{i+1}). \quad (5.43)$$

First note that

$$\mathbb{I}_{L,\ell} \sum_{i=L}^{L+\ell} \gamma_{i+1} M_{i+1} = \mathbb{I}_{L,\ell} \left(\sum_{i=0}^{L+\ell} \gamma_{i+1} \mathbb{I}_{i,0} M_{i+1} - \sum_{i=0}^{L-1} \gamma_{i+1} \mathbb{I}_{i,0} M_{i+1} \right). \quad (5.44)$$

By Lemma 5.8, $\{\mathbb{I}_{i,0} M_{i+1}\}_i$ is a martingale-increment. Therefore, by [16], a sufficient condition for $\sum_{i \geq 0} \gamma_{i+1} \mathbb{I}_{i,0} M_{i+1}$ to converge to zero is

$$\sum_{i \geq 0} \gamma_{i+1}^2 \mathbb{E} \left(\|\hat{H}_{\theta_i}(X_{i+1}) - P_{\theta_i} \hat{H}_{\theta_i}(X_i)\|^2 \mathbb{I}_{i,0} \right) < \infty. \quad (5.45)$$

By the parallelogram identity and Hölder's inequality,

$$\|\hat{H}_{\theta_i}(X_{i+1}) - P_{\theta_i} \hat{H}_{\theta_i}(X_i)\|^2 \mathbb{I}_{i,0} \leq 4 \sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_M} \|\hat{H}_\theta(x)\|^2.$$

Eqn. (5.45) then holds since $\sum_t \gamma_t^2 < \infty$. By (5.44), we obtain that

$$\lim_{L \rightarrow \infty} \sup_{\ell \geq 1} \left| \mathbb{I}_{L,\ell} \sum_{k=L}^{L+\ell} \gamma_{k+1} M_{k+1} \right| = 0 \quad \text{w.p. 1}.$$

Let us now consider the term $R_{i+1}^{(1)}$ defined in (5.42). Summing by parts, we get

$$\begin{aligned} \mathbb{I}_{L,\ell} \sum_{i=L}^{L+\ell} \gamma_{i+1} R_{i+1}^{(1)} &= \mathbb{I}_{L,\ell} \gamma_{L+1} P_{\theta_L} \hat{H}_{\theta_L}(X_L) + \mathbb{I}_{L,\ell} \sum_{i=L+1}^{L+\ell} (\gamma_{i+1} - \gamma_i) P_{\theta_i} \hat{H}_{\theta_i}(X_i) \\ &\quad - \mathbb{I}_{L,\ell} \gamma_{L+\ell+1} P_{\theta_{L+\ell+1}} \hat{H}_{\theta_{L+\ell+1}}(X_{L+\ell+1}). \end{aligned}$$

Since $\sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_M} \|\hat{H}_\theta(x)\| < \infty$, there exists a constant C such that the RHS is upper bounded by $C \left(|\gamma_{L+1}| + \sum_{i \geq L+1} |\gamma_{i+1} - \gamma_i| + |\gamma_{L+\ell+1}| \right)$. Under the stated assumptions, this upper bound yields

$$\lim_{L \rightarrow \infty} \sup_{\ell \geq 1} \left| \mathbb{I}_{L,\ell} \sum_{i=L}^{L+\ell} \gamma_{i+1} R_{i+1}^{(1)} \right| = 0,$$

with probability 1.

Finally, let us consider the term $R_{i+1}^{(2)}$ defined in (5.43). By (5.40), Lemma 5.12, and since on the event $\{\psi_{k+1} = \psi_k\}$, we have $\theta_{k+1} = \theta_k + \gamma_{k+1} H(X_{k+1}, \theta_k)$, we obtain

$$\begin{aligned} \mathbb{I}_{L,\ell} \left| \sum_{i=L}^{L+\ell} \gamma_{i+1} R_{i+1}^{(2)} \right| &\leq \mathbb{I}_{L,\ell} \sum_{i=L}^{L+\ell} \gamma_{i+1} \|P_{\theta_{i+1}} \hat{H}_{\theta_{i+1}} - P_{\theta_i} \hat{H}_{\theta_i}\|_\infty \\ &\leq C \mathbb{I}_{L,\ell} \sum_{i=L}^{L+\ell} \gamma_{i+1} \|\theta_{i+1} - \theta_i\|^{1-2\kappa} \leq C' \sum_{i=L}^{L+\ell} \gamma_{i+1}^{2(1-\kappa)}. \end{aligned}$$

This concludes the proof. \square

Lemma 5.14. *Let $M \in (0, M_\star)$ and set*

$$\Gamma_{M_\star}^M = \{\theta \in \Theta : M_\star \leq w(\theta) \leq M\}, \quad \iota = \inf_{\theta \in \Gamma_{M_\star}^M} |\langle \nabla w(\theta), h(\theta) \rangle|.$$

Under Assumptions 1 and 2, there exist $\delta \in (0, \iota)$ and $\lambda, \beta > 0$ such that

- (A) $u \in \mathcal{W}_{M_\star}, 0 \leq \gamma \leq \lambda, \|\xi\| \leq \beta \Rightarrow w(u + \gamma h(u) + \gamma \xi) \leq M$, and
- (B) $u \in \Gamma_{M_\star}^M, 0 \leq \gamma \leq \lambda, \|\xi\| \leq \beta \Rightarrow w(u + \gamma h(u) + \gamma \xi) < w(u) - \gamma \delta$.

Proof. Define $u' = u + \gamma h(u) + \gamma \xi$.

(A) Let $u \in \mathcal{W}_M$. Since w is continuous on Θ and the level set \mathcal{W}_M is a compact subset of Θ (see Lemma 5.6), there exists $\eta > 0$ such that for any $u \in \mathcal{W}_M$ and any u' satisfying $\|u' - u\| \leq \eta$, $u' \in \mathcal{W}_{M+1}$. Therefore, since

$$\|u - u'\| \leq \lambda(\max_{\mathcal{W}_M} \|h\| + \beta), \quad (5.46)$$

there exists $\lambda_1, \beta_1 > 0$ such that for any $0 \leq \gamma \leq \lambda_1$ and any $\|\xi\| \leq \beta_1$, $u' \in \mathcal{W}_{M+1}$ (note that $\max_{\mathcal{W}_M} \|h\| < \infty$ by Lemma 5.6).

Since w is continuous on the compact set \mathcal{W}_{M+1} (see Lemma 5.6), it is uniformly continuous (u.c.) on \mathcal{W}_{M+1} . Then we can choose $\lambda_2, \beta_2 > 0$ (smaller than λ_1, β_1) such that

$$\forall u \in \mathcal{W}_{M_*}, \forall \gamma \leq \lambda_2, \|\xi\| \leq \beta_2, \quad |w(u) - w(u + \gamma h(u) + \gamma \xi)| \leq M - M_* . \quad (5.47)$$

This concludes the proof of (A).

(B) Let $u \in \Gamma_{M_*}^M$. Following the same lines as in the proof of (5.47), there exist $\lambda_1, \beta_1 > 0$ such that for any $0 \leq \gamma \leq \lambda_1$ and $\|\xi\| \leq \beta_1$, $[u, u'] \subset \mathcal{W}_{M+1}$. By Lemma 5.6, this implies that w is continuously differentiable on (u, u') . We write

$$\begin{aligned} |\langle \nabla w(u), h(u) \rangle - \langle \nabla w(u'), h(u) + \xi \rangle| &= |\langle \nabla w(u), h(u) \rangle - \langle \nabla w(u'), h(u') \rangle \\ &\quad + \langle \nabla w(u'), h(u') - h(u) - \xi \rangle|. \end{aligned}$$

By Lemma 5.6, $\varphi : u \mapsto \langle \nabla w(u), h(u) \rangle$ is continuous and negative on the compact set $\Gamma_{M_*}^M$, so there exists $\epsilon \in (0, \iota)$ such that $\langle \nabla w(u), h(u) \rangle \leq -\epsilon$ on $\Gamma_{M_*}^M$. Furthermore, φ is u.c. on \mathcal{W}_{M+1} , and, for any $\epsilon' > 0$, we can thus take β_2 and λ_2 small enough so that for any $0 \leq \gamma \leq \lambda_2$ and $\|\xi\| \leq \beta_2$, $|\varphi(u) - \varphi(u')| \leq \epsilon'/2$. Therefore

$$|\langle \nabla w(u), h(u) \rangle - \langle \nabla w(u'), h(u) + \xi \rangle| \leq \epsilon'/2 + (\|h(u) - h(u')\| + \beta_2) \max_{\mathcal{W}_{M+1}} \|\nabla w\| .$$

Since $x \mapsto \|\nabla w(x)\|$ is continuous on the compact set \mathcal{W}_{M+1} , $\max_{\mathcal{W}_{M+1}} \|\nabla w\|$ is finite. As h is u.c. on \mathcal{W}_{M+1} , one can pick λ_2, β_2 small enough so that

$$\forall u \in \Gamma_{M_*}^M, \forall \gamma \leq \lambda_2, \|\xi\| \leq \beta_2, \text{ and } |\langle \nabla w(u), h(u) \rangle - \langle \nabla w(u'), h(u) + \xi \rangle| \leq \epsilon' .$$

Finally, applying Taylor's formula, we get

$$\begin{aligned} w(u') - w(u) &= \int_0^1 \left\langle \nabla w(u + t\gamma(h(u) + \xi)), \gamma(h(u) + \xi) \right\rangle dt \\ &= \gamma \varphi(u) + \gamma \int_0^1 \left(\langle \nabla w(u + t\gamma(h(u) + \xi)), h(u) + \xi \rangle - \langle \nabla w(u), h(u) \rangle \right) dt \\ &\leq -\gamma\epsilon + \gamma\epsilon' . \end{aligned}$$

Since ϵ' is arbitrary, this yields (B). \square

Proof of Item 1 in Theorem 3.2. Let $M > M_*$, let q (depending on M) be such that (see Remark 5.7)

$$\mathcal{W}_M \subset \mathcal{W}_{M+2} \subseteq \mathcal{K}_{\delta_q} , \quad (5.48)$$

and let $\theta_0 \in \mathcal{W}_M$. Let λ, β be given by Lemma 5.14. By Lemma 5.6, w and h are uniformly continuous on \mathcal{W}_{M+1} , and there exists $\eta > 0$ such that

$$x \in \mathcal{W}_M, \|x - y\| < \eta \implies |w(x) - w(y)| < 1 \text{ and } \|h(x) - h(y)\| < \beta . \quad (5.49)$$

By Lemma 5.13, there exists an almost surely finite r.v. N such that w.p.1.,

$$n \geq N \Rightarrow \gamma_n \left(1 + \sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_M} \|H(x, \theta)\| \right) < \lambda \wedge \eta, \text{ and} \quad (5.50)$$

$$\sup_{\ell \geq 1} \left(\prod_{i=N}^{N+\ell} \mathbb{1}_{\theta_i \in \mathcal{W}_{M+1}} \mathbb{1}_{\psi_{i+1}=\psi_i} \right) \left\| \sum_{i=N}^{N+\ell} \gamma_{i+1} (H(X_{i+1}, \theta_i) - h(\theta_i)) \right\| < \eta. \quad (5.51)$$

The proof is by contradiction. Denote by ψ_t the number of projections at the end of iteration t . We assume that $\mathbb{P}(\lim_t \psi_t = +\infty) > 0$. We can assume without loss of generality that

$$w(\theta_N) \leq M, \quad \psi_N \geq q$$

on the set $\{\lim_t \psi_t = +\infty\}$. Define the sequence $(\theta'_{N+k})_{k \geq 0}$ as

$$\theta'_N = \theta_N \quad \text{and} \quad \theta'_{N+k+1} = \theta'_{N+k} + \gamma_{N+k+1} h(\theta_{N+k}).$$

We prove by induction on k that for any $k \geq 0$, on the set $\{\lim_t \psi_t = +\infty\}$,

$$\theta'_{N+k} \in \mathcal{W}_M, \quad \theta_{N+k} \in \mathcal{W}_{M+1}, \quad \|\theta'_{N+k} - \theta_{N+k}\| < \eta, \quad \psi_{N+k+1} = \psi_{N+k}.$$

The case $k = 0$ is trivial since $\theta'_N = \theta_N \in \mathcal{W}_M$ and by using (5.49), (5.50), and (5.48) on the set $\{\lim_t \psi_t = +\infty\}$. Assume this property holds for $k \in \{0, 1, \dots, \ell\}$. Then we have

$$\theta'_{N+\ell+1} = \theta'_{N+\ell} + \gamma_{N+\ell+1} h(\theta'_{N+\ell}) + \gamma_{N+\ell+1} (h(\theta_{N+\ell}) - h(\theta'_{N+\ell})).$$

Since $\|\theta'_{N+\ell} - \theta_{N+\ell}\| < \eta$ and $\theta'_{N+\ell}$ is in \mathcal{W}_M , we have $\|h(\theta'_{N+\ell}) - h(\theta_{N+\ell})\| < \beta$. Since $\gamma_{N+\ell+1} < \lambda$ by (5.50), we can apply Lemma 5.14 to obtain $\theta'_{N+\ell+1} \in \mathcal{W}_M$. In addition,

$$\begin{aligned} \theta'_{N+\ell+1} - \theta_{N+\ell+1} &= \sum_{i=N}^{N+\ell} \gamma_{i+1} (H(X_{i+1}, \theta_i) - h(\theta_i)) \mathbb{1}_{\psi_{i+1}=\psi_i} + \sum_{i=N}^{N+\ell} (\gamma_{i+1} h(\theta_i) + \theta_i - \theta_0) \mathbb{1}_{\psi_{i+1} \neq \psi_i} \\ &= \left(\prod_{i=N}^{N+\ell} \mathbb{1}_{\theta_i \in \mathcal{W}_{M+1}} \right) \sum_{i=N}^{N+\ell} \gamma_{i+1} (H(X_{i+1}, \theta_i) - h(\theta_i)) \mathbb{1}_{\psi_{i+1}=\psi_i}, \end{aligned}$$

where we used the induction assumption in the last equality. From (5.49) and (5.51), this yields $\|\theta'_{N+\ell+1} - \theta_{N+\ell+1}\| < \eta$ and $w(\theta_{N+\ell+1}) \leq M + 1$. Finally by (5.49), Eqs. (5.50) and (5.48) imply that on the set $\{\lim_t \psi_t = +\infty\}$

$$\theta_{N+\ell} + \gamma_{N+\ell+1} H(X_{N+\ell+1}, \theta_{N+\ell}) \in \mathcal{W}_{M+2} \subset \mathcal{K}_{\psi_{N+\ell}},$$

that is, $\psi_{N+\ell+1} = \psi_{N+\ell}$. This concludes the induction.

As a consequence of this induction, we have $\psi_{N+\ell} = \psi_N$ for any $\ell \geq 0$ on the set $\{\lim_t \psi_t = +\infty\}$ which is a contradiction.

Proof of Item 2 in Theorem 3.2. The proof is along the same lines as the proof of Theorem 2.3 of [1, page 5], and is thus omitted.

5.7. Proof of Theorem 3.3

The proof consists in checking the conditions of [12, Corollary 2.8]. Let f be a measurable bounded function.

By Lemma 5.8, (i) there exists a measurable function \hat{f}_θ such that $\hat{f}_\theta - P_\theta \hat{f}_\theta = f - \pi_\theta f$; and (ii) for any compact set \mathcal{W}_M , there exists L (depending upon M) such that

$$\forall \theta \in \mathcal{W}_M, x \in \mathbb{X}, |\hat{f}_\theta(x)| \leq L.$$

By Theorem 3.2, $\mathbb{P}(\Omega_M) \uparrow 1$ when M tends to infinity where

$$\Omega_M = \bigcap_{t \geq 0} \{\theta_t \in \mathcal{W}_M\}.$$

Therefore, in order to apply [12, Corollary 2.8], we only have to prove that almost surely,

$$\sum_k k^{-1} \sup_{x \in \mathbb{X}} \|P_{\theta_k}(x, \cdot) - P_{\theta_{k-1}}(x, \cdot)\|_{\text{TV}} \mathbb{1}_{\Omega_M} < \infty, \quad (5.52)$$

$$\lim_t \pi_{\theta_t}(f) \mathbb{1}_{\Omega_M} = \pi_{\theta^*}(f) \mathbb{1}_{\Omega_M}. \quad (5.53)$$

By Lemma 5.11, there exists C and $\kappa \in (0, 1/2)$ such that

$$\sup_{x \in \mathbb{X}} \|P_{\theta_k}(x, \cdot) - P_{\theta_{k-1}}(x, \cdot)\|_{\text{TV}} \mathbb{1}_{\Omega_M} \leq C \|\theta_k - \theta_{k-1}\|^{1-2\kappa}.$$

In addition, by Theorem 3.2, there exists a random variable K , almost surely finite, such that for any $k \geq K$,

$$\|\theta_k - \theta_{k-1}\| \mathbb{1}_{\Omega_M} \leq \gamma_k \sup_{\theta \in \mathcal{W}_M, x \in \mathbb{X}} |H(x, \theta)|.$$

This yields

$$\sum_{k \geq K} k^{-1} \sup_{x \in \mathbb{X}} \|P_{\theta_k}(x, \cdot) - P_{\theta_{k-1}}(x, \cdot)\|_{\text{TV}} \mathbb{1}_{\Omega_M} \leq C \sum_{k \geq K} k^{-1} \gamma_k^{1-2\kappa},$$

for some constant $C > 0$. This concludes the proof of (5.52). The limit (5.53) is a consequence of Lemma 5.10.

5.8. Proof of Theorem 3.4

Let f be a measurable function such that $\|f\|_\infty \leq 1$ and set

$$I_t(f) = |\mathbb{E}[f(X_t) \mathbb{1}_B] - \pi_{\theta^*}(f) \mathbb{P}(B)| = |\mathbb{E}[(f(X_t) - \pi_{\theta^*}(f)) \mathbb{1}_B]|.$$

Let $\epsilon > 0$. We prove that there exists T_ϵ such that for all $t \geq T_\epsilon$, $\sup_{\{f: \|f\|_\infty \leq 1\}} I_t(f) \leq 4\epsilon$. Choose $\kappa \in (0, 1/2)$ and $\delta > 0$ such that

$$C_{M_\star+1} \delta^{1-2\kappa} \leq \epsilon, \quad (5.54)$$

where M_\star and C_{M_\star} are defined in Assumption 2 and in Lemma 5.10, respectively. Choose r_ϵ such that

$$2(1 - \rho_{M_\star+1})^{r_\epsilon} \leq \epsilon, \quad (5.55)$$

where $\rho_{M_\star+1}$ is defined in Lemma 5.8. By uniform continuity of w on $\mathcal{W}_{M_\star+2}$, assume finally δ is small enough that

$$\theta \in \mathcal{W}_{M_\star+1}, \theta' \in \Theta, \|\theta - \theta'\| \leq \delta \Rightarrow |w(\theta) - w(\theta')| \leq \frac{1}{r_\epsilon + 1}. \quad (5.56)$$

There exists T_ϵ^1 such that for any $t \geq T_\epsilon^1$,

$$\mathbb{P} \left(\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta, \lim_q \theta_q = \theta^\star \right) \leq \epsilon/2.$$

Hence, for any $t \geq T_\epsilon^1$, $I_t(f) \leq \sum_{i=1}^3 I_t^i(f) + \epsilon$, where

$$I_t^1(f) = |\mathbb{E}[(f(X_t) - P_{\theta_{t-r_\epsilon}}^{r_\epsilon} f(X_{t-r_\epsilon})) \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta}]| \quad (5.57)$$

$$I_t^2(f) = |\mathbb{E}[(P_{\theta_{t-r_\epsilon}}^{r_\epsilon} f(X_{t-r_\epsilon}) - \pi_{\theta_{t-r_\epsilon}}(f)) \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta}]| \quad (5.58)$$

$$I_t^3(f) = |\mathbb{E}[(\pi_{\theta_{t-r_\epsilon}}(f) - \pi_{\theta^\star}(f)) \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta}]|. \quad (5.59)$$

We first upper bound $I_t^1(f)$. For $\theta, \theta' \in \Theta$, let

$$D(\theta, \theta') = \sup_{x \in \mathbb{X}} \|P_\theta(x, \cdot) - P_{\theta'}(x, \cdot)\|_{\text{TV}}.$$

Applying [4, Proposition 1.3.1], it comes for any $t \geq T_\epsilon^1$,

$$\begin{aligned} I_t^1 &\leq \mathbb{E} \left[2 \wedge \sum_{j=1}^{r_\epsilon-1} D(\theta_{t-r_\epsilon+j}, \theta_{t-r_\epsilon}) \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta} \right] \\ &\leq \mathbb{E} \left[2 \wedge \sum_{j=1}^{r_\epsilon-1} (r_\epsilon - j) D(\theta_{t-r_\epsilon+j}, \theta_{t-r_\epsilon+j-1}) \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta} \right], \end{aligned}$$

where we used that for any $q, \ell > 0$ $D(\theta_{q+\ell}, \theta_q) \leq \sum_{j=1}^\ell D(\theta_{q+j}, \theta_{q+j-1})$. By Proposition 1, the random iteration number τ_ψ where the last projection occurs in Algorithm 2 is finite with probability one. Let then M_ϵ be such that $2\mathbb{P}(\tau_\psi \geq M_\epsilon) \leq \epsilon/2$, so that

$$I_t^1(f) \leq \mathbb{E} \left[2 \wedge \sum_{j=1}^{r_\epsilon-1} (r_\epsilon - j) D(\theta_{t-r_\epsilon+j}, \theta_{t-r_\epsilon+j-1}) \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta^\star\| \leq \delta} \mathbb{1}_{\tau_\psi \leq M_\epsilon} \right] + \frac{\epsilon}{2}.$$

Let now $T_\epsilon^2 \geq T_\epsilon^1 \vee (M_\epsilon + r_\epsilon)$ be such that

$$t \geq T_\epsilon^2 \Rightarrow \gamma_t \sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_{M_\star+2}} \|H(x, \theta)\| \leq \delta.$$

Then, by recurrence and using (5.56), we obtain that on $\{\|\theta_{t-r_\epsilon} - \theta_\star\| \leq \delta\}$, $\theta_{t-r_\epsilon+j} \in \mathcal{W}_{M_\star+1}$ for all $0 \leq j \leq r_\epsilon$. By Lemma 5.11 this yields for any $t \geq T_\epsilon^2$

$$I_t^1(f) \leq C_{M_\star+1} \left[\sup_{x \in \mathbb{X}, \theta \in \mathcal{W}_{M_\star+2}} \|H(x, \theta)\| \right]^{1-2\kappa} \sum_{j=1}^{r_\epsilon-1} (r_\epsilon - j) \gamma_{t-r_\epsilon+j}^{1-2\kappa} + \frac{\epsilon}{2},$$

and there exists $T_\epsilon^3 \geq T_\epsilon^2$ such that $t \geq T_\epsilon^3 \Rightarrow \sup_{\{f: \|f\|_\infty \leq 1\}} I_t^1(f) \leq \epsilon$.

We now consider $I_t^2(f)$; it holds

$$I_t^2 \leq \mathbb{E} \left[\left\| P_{\theta_{t-r_\epsilon}}^{r_\epsilon}(X_{t-r_\epsilon}, \cdot) - \pi_{\theta_{t-r_\epsilon}} \right\|_{\text{TV}} \mathbb{1}_{\|\theta_{t-r_\epsilon} - \theta_\star\| \leq \delta} \right].$$

By (5.56), $\|\theta_{t-r_\epsilon} - \theta_\star\| \leq \delta \Rightarrow \theta_{t-r_\epsilon} \in \mathcal{W}_{M_\star+1}$ and thus, applying Lemma 5.8 and (5.55)

$$\sup_{\{f: \|f\|_\infty \leq 1\}} I_t^2(f) \leq 2(1 - \rho_{M_\star+1})^{r_\epsilon} \leq \epsilon.$$

The derivation of the upper bound of I_t^3 is similar to that of I_t^2 , with Lemma 5.8 replaced by Lemma 5.10 and uses (5.54). Details are omitted.

Acknowledgements

This work was supported by the ANR-2010-COSI-002 grant of the French National Research Agency.

References

- [1] C. Andrieu, E. Moulines, and P. Priouret. Stability of stochastic approximation under verifiable conditions. *SIAM J. Control Optim.*, 44:283–312, 2005.
- [2] C. Andrieu and C. P. Robert. Controlled Markov chain Monte Carlo methods for optimal sampling. Technical Report 125, Cahiers du Ceremade, 2001.
- [3] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18:343–373, 2008.
- [4] Y. Atchadé, G. Fort, E. Moulines, and P. Priouret. *Bayesian Time Series Models*, chapter Adaptive Markov chain Monte Carlo: Theory and Methods, pages 33–53. Cambridge Univ. Press, 2011.
- [5] R. Bardenet, O. Cappé, G. Fort, and B. Kégl. Adaptive Metropolis with online relabeling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.
- [6] R. Bardenet and B. Kégl. An adaptive Monte Carlo Markov chain algorithm for inference from mixture signals. *J. Phys.: Conf. Ser.*, 368, 2012.
- [7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 1991.
- [8] J. W. Brewer. Kronecker products and matrix calculus in system theory. *IEEE Trans. Circuits Syst.*, 25:772–781, 1978.
- [9] G. Celeux. Bayesian inference for mixtures: The label-switching problem. In *Computational Statistics Symposium (COMPSTAT)*. Physica-Verlag, 1998.
- [10] G. Celeux, M. Hurn, and C.P. Robert. Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.*, 95:957–970, 1995.

- [11] A. J. Cron and M. West. Efficient classification-based relabeling in mixture models. *Amer. Statist.*, 65:16–20, 2011.
- [12] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2012.
- [13] S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Springer-Verlag, 2000.
- [14] P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- [15] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242, 2001.
- [16] P. Hall and C.C. Heyde. *Martingale limit theory and its application*. Academic Press, New York, 1980.
- [17] A. Jasra. *Bayesian inference for mixture models via Monte Carlo*. PhD thesis, Imperial College, London, UK, 2005.
- [18] A. Jasra, C. C. Holmes, and D. A. Stephens. Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statist. Sci.*, 20(1):50–67, 2005.
- [19] J.M. Marin, K. Mengersen, and C.P. Robert. Bayesian modelling and inference on mixtures of distributions. *Handbook of Statist.*, 25, 2004.
- [20] S.P. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Springer-Verlag, 1993.
- [21] G. Pagès. A space quantization method for numerical integration. *J. Comput. Appl. Math.*, 89:1–38, 1997.
- [22] P. Papastamoulis and G. Iliopoulos. An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distribution. *J. Comput. Graph. Statist.*, 19:313–331, 2010.
- [23] S. Richardson and P. J. Green. On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B*, 59(4):731–792, 1997.
- [24] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 2004.
- [25] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence of optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7:110–120, 1997.
- [26] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16:351–367, 2001.
- [27] G. O. Roberts and J. S. Rosenthal. Coupling and ergodicity of adaptive MCMC. *J. Appl. Probab.*, 44:486–475, 2007.
- [28] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *J. Comput. Graph. Statist.*, 18:349–367, 2009.
- [29] A. Roodaki. *Signal decompositions using trans-dimensional Bayesian methods*. PhD thesis, Supélec, Gif-sur-Yvette, France, 2012.
- [30] A. Roodaki, J. Bect, and G. Fleury. Summarizing posterior distributions in signal decomposition problems when the number of components is unknown. In *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, 2012.
- [31] M. Sperrin, T. Jaki, and E. Wit. Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statist. Comput.*, 20:357–366, 2010.
- [32] M. Stephens. Dealing with label switching in mixture models. *J. Roy. Statist. Soc. Ser. B*, 62:795–809, 2000.